DOCUMENT RESUME

ED 311 087 TM 013 944

AUTHOR	Rosser, Phyllis
TITLE	The SAT Gender Gap: Identifying the Causes.
INSTITUTION	Center for Women Policy Studies, Washington, D.C.
PUB DATE	89
NOTE	198p.
AVAILABLE FROM	Publications, Center for Women Policy Studies, 2000 P St., NW, Suite 508, Washington, DC 20036 ($15.00).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Ethnic Groups; Females; High Schools; High School Seniors; *Influences; Item Analysis; Males; *Predictive Validity; Racial Differences; *Sex Differences; Sex Discrimination; *Test Bias; Testing Problems; Test Items; Test Results
IDENTIFIERS	*Scholastic Aptitude Test

ABSTRACT

Questions on the Scholastic Aptitude Test (SAT) with the largest score differences between women and men of all racial and ethnic groups were identified. Patterns of difficulty that would explain the SAT's continuing underprediction of female first-year college performance were studied. An item analysis of one form of the June 1986 SAT for 1,112 students identified 17 questions with large sex differences. A subsequent item analysis of the November 1987 SAT used the responses of 100,000 high school seniors to identify 23 questions with substantial differences in the numbers of men and women who answered them correctly; 17 of these questions were in the mathematics section. African-American women exhibited the smallest gender gap and Hispanic women the largest when compared with men within their own racial and ethnic groups. The findings of this and other studies suggest that the test publisher could take steps to address the SAT's underprediction and bias. These are summarized as: (1) elimination of questions of identified bias; (2) testing a more balanced array of skills and knowledge; (3) publicity for validity studies about performance prediction; (4) more research on the correlation between SAT scores and grades; and (5) more test-taking time. Thirty-two supporting tables and nine appendices presenting aspects of the study are provided. A 219-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED381087

The SAT GENDER GAP

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

LESLIE R. WOLFE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Identifying the Causes

By Phyllis Rosser
CENTER FOR WOMEN POLICY STUDIES

BEST COPY AVAILABLE

The Center for Women Policy Studies was established in 1972 as the first independent national policy institute focused specifically on issues affecting the social, legal, and economic status of women. The Center's current programs, focused on issues affecting women and girls of color, include the National Resource Center on Women and AIDS, the Educational Equity Policy Studies Program, a policy analysis and seminar program on occupational segregation and its roots in education, the first National Conference of Young Women on issues of work and family, the second stage of the Reproductive Laws for the 1990s Project, and a Washington policy internship program for women of color. The Center's publications include influential materials on a range of issues, including: credit discrimination against women, sexual harassment of women in the workplace, women in poverty, rape, domestic violence, and equity in the Social Security system. The Center receives support from foundations, corporations, and individuals; the Educational Equity Policy Studies Program has received support from the Carnegie Corporation of New York, the Rockefeller Family Fund, the National Education Association, and others.

Typesetting and Design by Lexicon Graphics
Printing by Doyle Printing

The activity which resulted in this report was supported by a grant from the U.S. Department of Education, under the auspices of the Women's Educational Equity Act. Opinions expressed herein do not necessarily reflect the position or policy of the Department and no official endorsement should be inferred.

Discrimination Prohibited: No person in the United States shall, on the grounds of race, color or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance, or be so treated on the basis of sex under most educational programs or activities receiving federal assistance.

Additional copies of *The SAT Gender Gap* are available for $15.00 each from:
Publications
Center for Women Policy Studies
2000 P Street, NW Suite 508
Washington, DC 20036
202-872-1770

**(c) copyright Phyllis Rosser and Center for Women Policy Studies, 1989.
This publication may not be reproduced.**

The SAT GENDER GAP

Identifying the Causes

By Phyllis Rosser
CENTER FOR WOMEN POLICY STUDIES

Acknowledgements

I gratefully and humbly acknowledge the help of many people who have made this research possible. First thanks go to John Katzman for encouraging me to tackle the complexities of item analysis and for providing the Princeton Review students and computer work to make this possible. My next debt of gratitude goes to Jim Loewen at the University of Vermont who enthusiastically joined John and me and taught us the item analysis process. Thanks beyond measure, for the uncountable hours they spent with their computers, go to John for creating the computer programs and to Jim for analyzing the data and writing the results of our study, "Gender Bias in SAT Items" (in Chapter 2).

Immeasurable gratitude also goes to Leslie Wolfe, Executive Director of the Center for Women Policy Studies, for encouraging me to pursue this research and for the support she has given throughout the nearly two years of work it has entailed as well as for editing and publishing this report.

Another debt of gratitude goes to educational consultant Catherine Ross for her advice and analysis of the College Board magnetic tape data used in the study of gender bias on the November 1987 SAT (in Chapter 2). Thanks also are due to Philip McConnell at Boeing Commercial Airplane Company and to Jerry Julum at Boeing Computer Services in Seattle for accomplishing the challenging task of creating computer programs to extract the data from the College Board tape.

A bouquet of gratitude to Jackie Stevens who did a thorough and enlightening job of researching and summarizing the literature on sex bias in testing and the effects of test scores on girls' and women's educational opportunities. I also thank her for the moral support she continually gives me, reminding me that women are looking to my research to change their lives.

A round of thanks also goes to the following people, who have made my contribution to the sex bias in testing issue during the past ten years both possible and important: To the psychologists and educational researchers who taught me about testing and continue to educate and advise me—Mary Crawford, Patricia Campbell, Carol Tittle, Paula Selkow, Susan Klein, Susan Chipman, and Roger Chaffin. To John Weiss and the National Center for Fair and Open Testing—especially Bob Schaeffer, Sarah Stockwell, Denise Carty-Bennia, and Virginia Bullock—for publishing my earlier research, for making the issue of sex bias in testing known in households across the country, and for their continuing efforts to make tests fair for everyone. To Isabelle Pinzler and the staff of the Women's Rights Project of the American Civil Liberties Union and to Blair Horner and the New York Public Interest Research Group (NYPIRG), for their successful advocacy efforts in New York.

And finally, but most importantly, to Gloria Steinem and Paul Pottinger, who believed I could make a difference.

Phyllis Rosser
April, 1989

Table of Contents

Preface viii

Executive Summary 1

Chapter One—Defining Sex Bias in Standardized Testing 19

Introduction 21

The Impact of Sex-Biased Tests on Women's Educational Opportunities 22

The Underprediction of Women's Academic Performance by the SAT 23
- Underprediction by the ACT Assessment 26

Chapter Two—The SAT Gender Gap: Identifying the Causes 27

Gender Bias in SAT Items: An Initial Assessment 29
- Methods 29
- Results: Sex Differences 30
- The 17 Questions with Major Sex Differences 31
- The Students Most Affected by Sex-Biased Items 32
- Do Certain Types of Questions Favor One Sex? 33
- Do SAT Sex Differences Correlate with Performance Differences? 34
- Student Factors That May Cause Sex Differences in SAT Scores 36
- Socioeconomic Factors That May Cause Sex Differences in SAT Scores 39
- Effects of the SAT on Students 41
- SAT Differences, High School GPA Differences, and Perceived Ability, By Sex 41
- Do SAT Scores Influence Future Aspirations? 43
- Summary of Major Findings 44
- Implications for Test-Makers 45

Gender Bias on the November 1987 SAT: An Item Analysis 47
- Do Some SAT Questions Show Large Performance Differences By Sex and Race? 48
- The Questions with Major Gender Differences 50
  - Do Certain Types of Questions Favor One Sex? 54
  - The Math Score Gap 55
  - Do Women Choose Different Wrong Answers Than Men? 56
- Questions Showing Large Sex Differences Within Each Racial/Ethnic Group 57

- ☐ The "Racial/Ethnic Gap" 58
- ☐ Questions Showing Large Percentage Differences Between Women of Color and White Women 60
- ■ The Gender Gap at the Top: Correlating SAT Scores With High School Performance 60
- ■ Other Explanations 63
  - ☐ Omission of Questions 63
  - ☐ Time Pressure 64
  - ☐ Socioeconomic Factors: Parents' Education and Income 65

Is It Possible to Create a Sex-Fair Test? 68
- ☐ Construction of Sex-Biased and Sex-Equal Verbal Tests 68
- ☐ Construction of Sex-Biased and Sex-Equal Math Tests 68

High School Achievement Tests—Are They Fair for Girls? 71
- ☐ Girls' Score Averages Are Higher Than Boys' on the Major Standardized Achievement Tests Used in High School 72
- ☐ A Look at One High School's Experience 72
- ☐ Longitudinal Studies—Cause for Concern 73
- ☐ The Narrowing of Cognitive Differences 74

Review of the Literature on Gender Bias in College Entrance Examinations 75
- ■ Selected Studies 77

Chapter Three—Closing Doors: The Impact of Sex-Biased Tests on Women's Educational Opportunities 83
- ■ Using SAT Scores to Award Merit Scholarships 85
  - ☐ Sex Bias in National Merit Scholarship Awards 85
  - ☐ Using SAT Scores to Award State Merit Scholarships: State-by-State Analysis 85
  - ☐ Using SAT Scores to Choose "Gifted and Talented" Students: State-by-State Analysis 88
  - ☐ Private "Gifted and Talented" Programs—Exclusive Reliance on SAT Scores and Its Impact on Girls 88

College Admissions—Are SAT Scores Essential? 90
- ☐ Princeton University—A Case Study of Underprediction 92
- ☐ The Massachusetts Institute of Technology, Bates College and Bowdoin College—New Admissions Policies to Counter the SAT's Underprediction for Women 92

1988 Demographic and Selection Data for State Sponsored Summer Programs for Gifted High School Students 94

1988 Survey of Statewide Merit Scholarship Programs 107

Chapter Four—Recommendations for Further Research and Development 117
- ■ Recommendations for Test Publishers 119
- ■ Recommendations for Further Research 120

Bibliography and References 123

Appendices 139

List of Tables

Table 1. SAT Averages by Sex 31

Table 2. The 7 SAT Verbal Items That Favored One Sex by Approximately 10 Percent or More 31

Table 3. The 10 SAT Math Items That Favored One Sex by More Than 10 Percent 32

Table 4. Mean Differences by Sex in Percentage Correct on Sex-Biased Items, Among Low, Middle, and High SAT Scorers 33

Table 5. Scores by Sex on Different Types of Items 34

Table 6. Percentage Reporting Various High School GPAs, by Sex 35

Table 7. Percentage Reporting Various High School GPAs, by SAT Score Range and Sex 35

Table 8. "How Do You Feel About the SAT?" by Sex 36

Table 9. Math SAT Items Correct by Math as Favorite Subject, by Sex 37

Table 10. Math Preparation by Sex 38

Table 11. Math SAT Items Correct by Amount of Math Taken, by Sex 39

Table 12. Mean Number of SAT Items Correct Correlated to Mother's Occupation, by Sex 40

Table 13. Ratio of Girl/Boy Ranking on English High School GPA, Verbal SAT Scores, and Perceived Verbal Abilities 42

Table 14. Ratio of Girl/Boy Ranking on Math High School GPA, Math SAT Scores, and Perceived Math Abilities 42

Table 15. Percent of Students Who Place Themselves In the Listed Percentile Groups in Self-Perceived Abilities, as Affected by SAT Scores 43

Table 16. Students Who Plan to Attend Different Types of Colleges, By High School GPA 44

Table 17. Students Who Plan to Attend Different Types of Colleges, By SAT Scores 45

Table 18. SAT Averages by Sex 48

Table 19. Score Averages by Sex and Race, With Female Difference Within Racial/Ethnic Group [November 1987 Sample] 49

Table 20. Ranking of Each Racial/Ethnic Group by Combined Scores, from Highest to Lowest, [November 1987 Sample] 50

Table 21. Average SAT Scores for Each Racial Group, Highest to Lowest, for November, 1987 Sample 51

Table 22. 6 SAT VERBAL Items Favoring One Sex by Approximately 10 Percent 51

Table 23. 17 SAT MATH Items Favoring One Sex by More Than 10 Percent or a Large Ratio 53

Table 24. Easy, Medium, Difficult Items 54

Table 25. Scores by Sex on Different Types of Items 55

Table 26: Items With Wide Male/Female Variance, by Race/Ethnicity Showing Percentage Differences, and Ratios 59

Table 27. Average SAT Scores for Females and Males in Each GPA Category 62

Table 28. Questions That Were More Difficult for Girls Than Boys with A+ GPAs 62

Table 29. Comparison of GPA to SAT Quartiles by Sex 63

Table 30. November 1987 SAT Results With No Questions Removed 70

Table 31. November 1987 SAT Results With Four "Worst" Questions Removed From Each Section [Raw Scores Only] 70

Table 32. Correlations Between Grades and Test Scores for 203 High School Seniors 73

Appendices

Appendix A
Items With Extreme Differences by Sex (June 1986 SAT) 141

Appendix B
Questionnaire Used With Princeton Review Students (June 1986 SAT) 145

Appendix C
Technical Notes for Study of Gender Bias in the June 1986 SAT 149

Appendix D
Items with Extreme Differences by Sex (November 1987 SAT) 153

Appendix E
Distractors Chosen for Questions That Create the Largest Sex Differences (November 1987 SAT) 159

Appendix F
Percentage Differences for Women of Color Compared to White Women for All Questions (November 1987 SAT) 162

Appendix G
Number of Omissions for Each Question, by Sex (November 1987 SAT) 167

Appendix H
Females' Average Scores Are Lower Than Males' At Each Income Level 171

Appendix I
Opinion and Order by Judge John M. Walker (United States District Court, Southern District of New York) in the case of Khadijah Sharif, *et. al.* against New York State Education Department *et. al.* 173

Preface

"After a careful review of the evidence, this Court concludes that SAT scores capture a student's academic achievement no more than a student's yearbook photograph captures the full range of her experiences in high school." With this vivid statement, Federal District Judge John M. Walker crystallized a concern that has occupied Phyllis Rosser for a decade. In February of 1989, in a case brought by the ACLU Women's Rights Project on behalf of high school women in New York State, Judge Walker ruled that the exclusive use of SAT scores to award merit scholarships to New York high school students discriminates against girls (*New York Times*, February 4, 1989; see Appendix I for the complete Opinion and Order).

This ruling was a milestone in a controversy that began in December of 1985, when Phyllis Rosser's important article, "Do SATs Shortchange Women?," was published in *Ms.* Magazine. Although bias has existed in the SAT since its initial publication in 1926 and had been reported in the research literature for several years, the test was not widely considered unfair to women until recently and Rosser's article was virtually the first in the popular press to report on this research. Thanks to Rosser's determination to bring these complex issues to public attention, many women learned from this article that they had earned higher average grades than men in both high school and college but had received lower average SAT scores—by a "gender gap" of approximately 60 points. Indeed, Rosser's investigative work had found that this important college entrance examination, published by the Educational Testing Service (ETS) and designed to predict first year college grades, has consistently underpredicted women's academic achievement during the past 22 years.

In December of 1986, in collaboration with the National Center for Fair and Open Testing (FairTest), Rosser produced another first at the FairTest Washington conference; she convened a panel of scholars and advocates to discuss the nature and extent of sex bias in standardized testing and to develop recommendations for further research and policy development. During the last two years, this work has progressed rapidly, culminating in the publication of *The SAT Gender Gap—Identifying the Causes* by the Center for Women Policy Studies.

In 1987, Rosser conducted research on the impact of sex-biased tests on young women's educational opportunities for FairTest; she discovered that girls received approximately one-third of the National Merit Scholarships, while boys received two-thirds—because girls received lower average scores on ETS's Preliminary Scholastic Aptitude Test/National Merit Qualifying

Test (PSAT/NMQT)—which is used as the sole criterion to qualify for these prestigious scholarships. Rosser's results, published by FairTest in April of 1987 in *Sex Bias in College Admission Tests: Why Women Lose Out*, and her testimony before the Subcommittee on Civil and Constitutional Rights, chaired by Representative Don Edwards (D-CA), have generated increasing coverage of the gender gap in the awarding of National Merit Scholarships and have focused national attention on sex bias in college entrance examinations generally.

Reporters, advocates for educational equity for women and girls, and others concerned about the use of standardized tests to evaluate student performance and capability have eagerly asked for examples of biased questions. But in 1987, only the test publisher knew which questions showed marked differences between male and female test takers and these data were not made publically available. The *SAT Gender Gap—Identifying the Causes* remedies that situation.

With funding from the Women's Educational Equity Act Program in the U.S. Department of Education, Phyllis Rosser conducted two major item analyses of the SAT, looking at the percentage of correct answers for men and women on every question on two tests—the June 1986 SAT and the November 1987 SAT. The purpose of this study was to identify those questions with the largest score differences between women and men of all racial/ethnic groups and to ascertain whether there are patterns of difficulty that would explain the SAT's continuing underprediction of female academic performance.

Thus, Phyllis Rosser is one of the few researchers outside of ETS who has identified questions that are considerably more difficult for girls of all racial/ethnic groups. These questions are published here in hopes that this effort will inspire and facilitate more research on this important barrier to educational equity for women and girls.

The Center for Women Policy Studies is pleased to publish *The SAT Gender Gap* as an initial product of our continuing research, policy development, and advocacy work on the nature and impact of sex and race bias in standardized testing.

Leslie R. Wolfe
Executive Director
Center for Women Policy Studies
April, 1989

Executive Summary

Executive Summary

Chapter 1—Defining Sex Bias in Standardized Testing: Standardized tests are widely used as achievement tests in elementary and secondary schools to evaluate academic progress and to identify students in need of compensatory education. They are also used as aptitude tests in the college admissions process to predict a prospective student's first year grades. Sex bias can be expressed in four ways: in *test content*, when many more men than women are referred to or depicted and women are shown in lower status or stereotyped roles; in *test context*, when questions are set in experiences more familiar to one sex than the other; in *test validity*, when women's academic abilities are underpredicted by test scores while men's are overpredicted; and, in *test use*, when women's access to educational opportunities is diminished or restricted by an institution's reliance on a test that underpredicts their abilities.

The form of sex bias that has the greatest negative impact on women's educational opportunities is the underprediction of their first year college performance by both of the major college admissions tests—the SAT, published by Educational Testing Service (ETS) and taken annually by 1.5 million students (52 percent of whom are female), and the American College Testing Program's ACT Assessment, taken annually by nearly one million students (54 percent of whom are female).

The major purpose of these tests is to predict first year college grades. But studies show that women earn higher average grades than men in all subjects in both high school and college classes from their first year onward. Yet, women receive lower average scores than men on both the SAT and the ACT. They also receive lower average scores on the Preliminary Scholastic Aptitude Test/National Merit Qualifying Test (PSAT/NMQT), published by ETS and taken annually by approximately 1.1 million high school juniors, 54 percent of whom are female. Although the PSAT is defined as a "practice test" for the SAT, the National Merit Scholarship Corporation awards over $23 million in scholarships each year to the students with the highest scores on this test, making it extremely important as a "gateway" to college for many students.

■ **The Impact of Sex-Biased Tests on Women's Educational Opportunities:** Reliance on these biased tests has an adverse impact on young women's educational opportunities in three important ways. By underpredicting their academic performance, the tests affect women's chances to gain entrance to nearly 1500 four-year colleges and universities that require SAT scores or use SAT cut-off scores for admission. Unfairly low test scores also become a self-fulfilling prophecy, causing young women

to lower their expectations and apply to less competitive colleges and universities than their grades would warrant. Lower test scores also exclude secondary school girls from academic enrichment programs and accelerated courses, including summer programs for "gifted and talented" students who are defined initially as those 7th through 11th graders who score 430 or higher on the Verbal Section of the SAT and 500 or higher on the Math Section (on a scale of 200-800).

Reliance on biased tests has a severe economic impact on women, who lose millions of dollars in merit scholarship awards—which are awarded annually by 22 states as well as hundreds of corporations, foundations, professional organizations, unions and government agencies—based on SAT, ACT or PSAT scores. The National Merit Scholarship Corporation, which offers the most prestigious awards for academic excellence, selects its semifinalists *solely* on the basis of PSAT scores. In 1987-88, women's PSAT scores averaged 54 points lower than men's and their qualifying scores (the verbal score doubled with the math score added) were 67 points lower, leaving women eligible for only 36 percent of the approximately 6,000 scholarships.

■ **The Underprediction of Women's Academic Performance by the SAT:** In 1988, women's average SAT scores were 56 points lower than men's: 13 points on the Verbal Section—where women excelled until 1972 when men began to outscore them—and 43 points on the Math Section. However, the College Board's own Validity Studies show that women's average first year college grades are as good or better, in all subjects, than are those of their male peers.

Therefore, the SAT does not fulfill its primary purpose—the prediction of first year college performance—for women. If the SAT *were* predictive, these young women would either earn lower first year college grade point averages than they actually do or they would receive higher average test scores, perhaps 10 or 20 points higher than men rather than 56 points lower. Since 52 percent of the test taking population is female, this test is underpredicting grades for approximately 780,000 young women every year.

The College Board, which administers the SAT, reported in 1988 that women who took the test had a far higher mean Grade Point Average (GPA) than the men who took the test. Of students with the highest grades (A+), 53 percent were women and 47 percent were men; women were 58 percent of A students and 54 percent of B students. While SAT scores for both men and women declined from 1973 to 1982, high school grade point average and class rank have remained consistently higher for women than for men.

The gender gap also cannot be attributed to large variations in academic preparation. The College Board reports that in 1988, 88 percent of the women had taken four years of English compared to 86 percent of the men; 97 percent of both sexes had taken algebra and 93 percent of the males and 92 percent of the females had taken geometry, reportedly all the math needed for the SAT. The SAT's underprediction for women has not been a secret. ETS researchers Clark and Grandy (1984) state that "the underprediction of women's first year college grades has been reported consistently in the research literature" (p. 21).

■ **Race-Plus-Sex Bias—The Impact on Women of Color:** Women of color are doubly penalized by the SAT. They all score lower than the men in their racial/ethnic group, according to the latest College Board Report (1988). All men of color, in turn, receive lower combined average scores than white men. For example, African American women averaged 32 points lower than African American men in 1988 and 241 points lower than white men;

African-American men averaged 209 points lower than white men.

■ **Students with Disabilities and the SAT:** Approximately 6,000 "nonstandard" SATs are administered to students with disabilities each year. Although the literature review conducted for this study did not find any studies that compared male and female differences, limited research has been done on the testing of students with visual, hearing, and physical impairments and students with learning disabilities. These studies show the SAT is generally less predictive for students with disabilities.

■ **Underprediction for Women by the ACT Assessment:** Women's college performance is also being underpredicted by the ACT Assessment, the other major college entrance examination, which is taken by nearly a million students in the Midwest, Southwest and South. In 1987-88, the average ACT Composite Score for men was 19.9 compared to 18.6 for women. Researchers have found that all ACT subject scores and the ACT Composite score (the average of the combined subject scores) "consistently underpredicted" women's two-year cumulative college Grade Point Average, even when partially controlled for different courses taken. The ACT is also having an adverse impact on male and female students of color, who all receive lower ACT composite scores than white males. And women within each ethnic group receive lower scores than men.

Chapter 2: The SAT Gender Gap—Identifying the Causes: To determine whether individual questions were creating the gender gap, two item analyses (an examination of responses to each question) were conducted. An initial study of 1,112 coaching students (conducted by James Loewen, Phyllis Rosser, and John Katzman) served as a preliminary study for a larger item analysis (conducted by Rosser) of 100,000 students who took the November 1987 SAT.

Men have always received higher scores than women on the SAT since its first administration in 1926, but their higher math scores were once partly offset by women's higher verbal scores (by approximately 5 points). Women lost their verbal lead in 1972, due to gradual changes in the test content that added questions referring to science, business, and "practical affairs" and eliminated questions with human relations, arts, and humanities content. According to ETS researchers, the test was changed to create "a better balance for the scores between the sexes." As a result, by 1986, the verbal gender gap favored men by 11 points. Although this change in test specifications required more male-oriented items on verbal tests, where women traditionally excel, the reverse (more female-oriented items on math tests, where men traditionally excel) has not been required; this has been called "nonconscious sexism" by an ETS researcher.

■ **Gender Bias in SAT Items: An Initial Assessment:** An item analysis was conducted of one form of the June 1986 SAT, to determine whether specific questions or other factors were creating or widening the score gap between the sexes and to determine how SAT scores influenced students' future academic plans. In March, 1987, 1,112 students in Princeton Review coaching classes took one form of the June 1986 SAT along with a 25-item questionnaire (Appendix B), which asked them to indicate their high school grade point averages (GPA), favorite high-school subjects, perceived ability in English and math, test anxiety, and family background. All students came from New York City high schools; 55.6 percent were girls and 44.4 percent were boys (nationally, SAT takers are 52 percent female); 75.3 percent were white, 13.2 percent Asian Americans, 5.2 percent African Americans, and 2.4 percent Hispanics. Almost all (97.8 percent) were in the 11th grade and

57 percent reported grade point averages from B+ to A+. Their high school preparations were strong: 86 percent had taken three years of math and 92 percent of the girls and 91 percent of the boys had taken three years of English in their three years of high school. Most students came from upper-middle class backgrounds; 81 percent of their fathers and 52 percent of their mothers pursued professional careers (doctors, executives, engineers, teachers, for example) and 72 percent of their fathers and 63 percent of their mothers were college graduates. Although this sample cannot be seen as random or representative of the national population, their uniformity in socioeconomic status is especially valuable as it allows an exploration of differences by sex that cannot be attributed to low incomes or lesser educational preparation.

■ **Results—17 Questions With Major Sex Differences:** On the Verbal SAT nationally, men now outscore women by about 10 points; but in this sample, males and females scored equally well. On the Math test, men outscore women nationally by about 47 points; in this sample, males outscored females by about 35 scale points. Girls and boys scored within a few percentage points of each other on most verbal and math questions, reflecting the fact that wide areas of experience, skills and sub-cultural terms are shared by young people of both sexes, and that most SAT questions tap those areas. However, 7 of the 85 verbal and 10 of the 60 math items showed considerable differences (more than 10 percent) in the percentage of each sex that answered them correctly.

Thirteen questions favored boys and 4 favored girls (see Appendix A for the full text of these questions). The 7 verbal items with large gender differences reflect traditional sex stereotypes; for example, words referring to relationships ("requite"), jewelry ("pendant"), and fabric ("sheen") favored girls while items such as the analogy "mercenary is to soldier as hack is to writer" favored boys.

Among math items, 10 differences of greater than 10 percent appeared, all favoring men. Three of these math items were specifically about boys' enterprises, suggesting that verbal bias adversely affects girls' performance on math items; the question with the largest gender difference (27 percent) required computation of a basketball team's won/loss record. (Earlier studies have shown that when math content is made relevant to female experience, males do not outperform females on math problems.)

This study confirmed the underprediction that other researchers have noted: girls received lower average scores than boys on the SAT, yet they earned higher average high school grades than boys in both English and math. The study also found significant item bias, suggesting that ETS's review process is less successful than it should be and that biased questions contribute to the gender gap on the SAT. Specific item content made the greatest difference, rather than type of item, academic subject matter, or level of difficulty.

The study also found that girls' poorer performance was not linked to test anxiety or time pressure, which are often postulated as reasons for women's lower scores. While boys liked math somewhat better and took slightly more math, this only explained part of their SAT-Math lead over girls. Controlling for social class still produced a score gap favoring boys. Finally, when estimating their math and English abilities, both men and women perceived their abilities to be more in line with their test scores than with their grades. Unfortunately, this meant that girls saw themselves as less able than their grades would indicate, and less able than boys.

These findings about young women's self-perceptions and aspirations remain troubling. Although girls and boys earned almost identical grades in math, only 38 percent of girls put themselves in the top 10 percent in math ability, compared to 56 percent of boys, confirming earlier studies that found that students' overall perceptions are closer to test feedback than to grade feedback. While this may be beneficial for boys' self image, it is quite damaging to girls,' because they tend to internalize the SAT's underprediction of their academic performance as an assessment of their "aptitude." Young women have a lower perception of their math ability even when they do well on the Math SAT. The study found that 57 percent of high-scoring boys put themselves into the top 5 percent in math ability, while only 39 percent of the girls did so. Even when the test tells them they are "good at math," girls are less likely to believe it.

■ **Gender Bias on the November 1987 SAT—An Item Analysis:** This item analysis is based on the responses of 100,000 college-bound high school seniors to one form of the November 1987 SAT, contained on a College Board data tape compiled by ETS. The sample represented nearly all the students who took one of the four forms of the test administered at that time and is the best random sampling of the student population that ETS makes available to the public.

The results of this item analysis represent a substantial new body of data to explain the causes of the gender gap in SAT scores. This research is among the first by an independent researcher, not affiliated with ETS, that uses ETS data in its attempt to determine whether specific questions create or contribute to the score gap, whether the SAT correlates with current academic performance for both sexes, and whether other factors might be causing sex differences.

■ **Do Some SAT Questions Show Large Performance Differences by Sex and Race?:** Women received lower average scores than men on both sections of the SAT—14 points lower on the Verbal Section and 44 points lower on the Math. And women in every ethnic group received lower average scores than the men in their ethnic group. The largest score gap occurred between Hispanic women and men (69 points) and the smallest between Asian American women and men (48 points). Although white males received the highest average scores (974) and African American females the lowest (759), Asian American males averaged 26 points higher than white males on the SAT-Math. Asian American females averaged only 14 points lower than white males on the SAT-Math, in contrast to white females, who averaged 43 points lower. This finding raises interesting questions about potential differences in the preparation of girls of different racial/ethnic backgrounds in mathematics.

■ **The 23 Questions with Major Gender Differences:** Of the 145 questions on the test, 23 (16 percent) displayed substantial differences in the number of women and men who answered them correctly. A closer analysis was conducted of all questions with an approximately 10 percent or greater difference between females and males in the percentage of correct answers or a large difference in the proportion (ratio) of females to males who answered them correctly. In the Verbal Section, girls scored considerably lower than boys on 4 questions and higher on 2 questions; for the full text of all 6 questions see Appendix D. A larger percentage of women than men chose the correct answers for questions referring to relationships and a larger percentage of men chose the correct answers for questions referring to physical science, sports, and the stock market.

Among the 60 Math questions, 17 exhibited large (10 percent or more) percentage or ratio differences between the sexes, *all* favoring men, who outscored women on every math question on this test, despite their lower average math grades. The pattern in math word problems is worth noting, as young women found 6 of the 10 word problems on the test considerably more difficult than did their male peers, regardless of item content.

■ **Do Certain Types of Questions Favor One Sex?:** This study found that girls performed slightly better than boys on the easy Verbal items and somewhat worse on the difficult items but the difference was not large. Unlike the initial study, large gender differences did appear in comparing the "easy," "medium" and "difficult" questions in the Math Sections.

Earlier studies have found that women perform better on reading comprehension questions and antonyms and worse on analogies and sentence completion questions but this item analysis found that girls performed slightly worse on all types of questions. Past research has found that girls perform less well in geometry than in algebra or arithmetic and the findings of this item analysis confirm this; the male average percentage correct for geometry questions was 8.8 percent higher than the females'. Arithmetic questions showed the smallest math difference between the sexes; the male average percentage correct was 4.86 percent higher than the females'. However, earlier research found that SAT arithmetic items favored girls, so this raises the question of what is causing this change.

■ **The Math Score Gap:** The mathematical score gap between the sexes has been present on the SAT at least since 1967, when the College Board first published national data on college-bound seniors. Apparently it has always existed but "efforts have not been made to 'balance' the SAT quantitative sections, even though sex differences have favored males by a great number of points since the first administrations of the test," according to an ETS researcher.

A recent study by Gross and Sharp of more than 4,000 high school students in Montgomery County, Maryland public schools, found that girls who took the same advanced math courses as boys—calculus, pre-calculus and advanced algebra—in the same classrooms and with the same teachers, earned higher grades but received SAT-Math scores that were 37 to 47 points lower than the boys.' Kanarek's study of Rutgers University's class of 1985 first year students (which included more than 1,000 women) found that the women had higher average grade point averages than the men in science and math; their GPAs in the humanities were substantially higher than the men's.

The fact that female performance on the Math Section of the SAT has always been worse than males', despite women's higher math grades, and that "balance" has not been attempted or achieved, raises important questions about the intent of the test publishers. Test questions are written to meet the publisher's content specifications; what decisions have ETS test developers made to justify the lack of prediction on this section of the test?

■ **Questions Showing Large Sex Differences Within Each Racial/Ethnic Group:** African American women exhibit the smallest gender gap and Hispanic women the largest, when compared to men within their own racial/ethnic group. This study sought to determine which questions were creating the problems and whether there was a discernible pattern. Only one verbal question made a large difference for women in every racial group—a Sentence Completion question set in a sports context (Appendix D). However, the rest of the questions that created a gender gap within racial/ethnic groups did not form general patterns that could be analyzed;

they are listed in Chapter 2 of this report and compared in Table 26. This research is the first to make questions creating gender differences within racial/ethnic groups available to the public and is intended to inspire further research.

A total of 38 math questions created a gender gap for women of color. The mathematics gender gap was smallest for African American women; although they scored lower than any other ethnic/gender group on the test, there were only six math questions with differences of more than 10 percent or large ratio differences compared to African American men. Native American women had the largest math gender gap, with 24 questions that had substantial differences. Hispanic women followed with 22 questions, white women with 18 and Asian American women with 16.

■ **The "Racial/Ethnic Gap":** Virtually no prior research has been published on the differences between female and male performance in any racial/ethnic group other than African Americans, nor are comparisons usually made across the racial/ethnic spectrum (comparing men and women of color to white men and women). Perhaps this lack of research is due to the fact that the gender difference within racial/ethnic groups is so much smaller than the well documented gap between white students and students of color. The outstanding exception has been the math performance of Asian Americans; men outperform, and females score almost as well as, white males. Studies of the "racial/ethnic gap" are reviewed in Chapter 2.

■ **Questions Showing Large Percentage Differences Between Women of Color and White Women:** African American women in this study performed worse than white women on every question on the test. Over half the Verbal questions (53 out of 85) and 80 percent of the Math questions (49 out of 60) showed differences of more than 10 percent or had large ratio differences. An even greater difference was found in comparing both groups to white men (white women averaged 57 points lower than white men). African American women (who averaged 241 points lower) performed worse on every question, compared to white men, with 71 percent (60 out of 85) of the Verbal and 82 percent of the Math questions showing differences of more than 10 percent.

Hispanic women performed better than white women on 5 of the Verbal questions. On one question Hispanic women performed more than 10 percent better ("the opposite of 'commodious'"), but they were more than 10 percent lower in correct answers or had large ratio differences on almost half the Verbal questions (42 out of 85) and over two-thirds of the Math questions (43 out of 60). Native American women found one question considerably easier than did white women ("Rebel:Insurrection"), but they did much worse than white women on 20 Verbal questions and 28 Math questions. On the other hand, Asian American women performed better than white women on 80 percent of the Math questions; they scored somewhat higher on 42 questions and more than 10 percent higher on 6 questions. They did better on 8 Verbal questions but worse on 24 others.

These data suggest that, with the exception of Asian American women, a large number of questions are causing the score differences between women of color and white women and white men. Appendix F includes all of the questions which had a 10 percent or greater difference in correct answers or a large ratio difference for women of color.

■ **The Gender Gap at the Top: Correlating SAT Scores with High School Performance:** It was surprising and distressing to find, in comparing high school grades to SAT scores, that *the higher the grades, the larger the gender gap.* The

biggest sex differences in SAT score averages—much larger than the national averages for the test as a whole—occurred at the highest GPA level (A+ to A), while the smallest gender gap occurred at the lowest GPA level. Women with A+ grades averaged 23 points lower on the Verbal Section than men with A+ grades; this is a substantially larger gap than for women in general (14 points). Further, these A+ women scored 60 points lower than A+ men on the Math Section, compared to 44 points for women in general.

A College Board representative has explained the larger math gap by claiming that women with A+ grade point averages are more likely to have earned them in English, humanities and language courses while the A+ men are more likely to have taken courses that prepared them for the SAT-Math, such as physics, chemistry and calculus. However, this fails to account for the larger gender gap on the SAT-Verbal Section, where one would expect the high achieving girls with English and humanities backgrounds to excel.

This is one of the most important findings of this study—that the highest achieving girls are penalized most by the SAT score gap. Their lower SAT scores in comparison to high achieving boys make the test less predictive for them. This may exclude them from the most prestigious colleges that accept their male peers and may also prevent them from qualifying for merit scholarships and other scholarships that are based on SAT scores rather than high school performance.

■ **Other Explanations: Omission of Questions:** Another critical discovery came from the analysis of the number of women and men who omitted each question (left the answer blank) on the test: a larger percentage of girls than boys left 50 of the 60 math questions blank. An even larger percentage of girls omitted the last 5 questions in both Verbal Sections and the last 10 questions (except one) in both Math Sections (the number of omissions for each question, by sex, can be found in Appendix G). Several theories suggest explanations for girls' greater tendency to omit items.

Some research shows that girls are less likely to be risk-takers and to guess at the right answer, largely because of their different upbringing, socialization, and earlier education. Linn *et. al.* found that 13 to 17 year old girls were more likely to use the "I Don't Know" response on the National Assessment of Educational Progress (NAEP) science assessment, "especially for items with physical science content or masculine themes such as football." Research on NAEP math tests also has found that gender differences appeared *favoring females* when the "I don't know" option was removed. These test results correlated well with the students' 7th and 10th grade classroom performance, where girls were earning higher grades than boys, in contrast to NAEP tests with the "I don't know" option, where girls scored worse than boys.

Another conclusion that could be drawn from these studies is that girls may be more likely to follow instructions or "play by the rules." Before each administration of the SAT, the monitor tells students that 1/4 point is subtracted from their score for each wrong answer but nothing is subtracted if the question is left blank. This warning about the "guessing penalty" is probably taken more seriously by girls (the "guessing penalty" has been removed from the Graduate Record Examination (GRE) but not from the SAT). As Harvard's Carol Gilligan told Rosser in 1987, "this test is a moral issue for girls; they think it is an indication of their intelligence, so they must not cheat. But boys play it like a pinball game."

■ **Time Pressure:** Males' and females' performance on the last 10 items on each section of the test—where they might run out of time—were compared to their performance on the rest of the test and to each other. Although a

larger percentage of girls than boys omitted the last Verbal questions, large percentages of both boys and girls omitted questions in the middle of the test. In a number of cases, larger percentages of boys than girls omitted these questions, indicating that content as well as timing was a problem for both sexes.

However, the omissions on the Math Sections told a different story. Large percentages of both males and females omitted the last 10 questions on the Math Sections, compared to their omissions on the rest of the sections, indicating that *both* boys and girls ran out of time. But on most questions, a much larger percentage of girls than boys omitted them, indicating that girls have a greater problem with time pressure on the Math Sections of the test than boys do.

It is important to note that the artificial emphasis on speed in the SAT is the antithesis of the current educational interest in teaching higher level thinking skills. This highly speeded test rewards the facile test taker rather than the sophisticated, thoughtful thinker who gathers new information and organizes, evaluates, and expresses original thoughts clearly and concisely.

■ **Socio-Economic Factors:** While this study corroborated other research which has found that social class, measured by parental education and income, was highly correlated with SAT performance for both sexes, it also found a significant gender gap at the highest socioeconomic level.

■ **Parents' Education:** The 100,000 students in the sample were separated into six levels of parental education. Comparing the percentage of correct answers for females and males in each level showed—surprisingly—that higher levels of parental education did *not* narrow the gender gap.

■ **Parental Income**: The most unexpected finding in this socioeconomic-status cluster came from comparing girls and boys from high income homes. Although SAT scores rise with family income level, there is still a high income gender gap—girls from the highest income families (over $70,000) receive lower average scores than boys at this income level. In fact, their Math score averages are the same as those of boys from the middle income range ($40-50,000).

This significant finding indicates that class does not predict SAT scores for girls the way it does for boys. When ETS suggests that the larger numbers of low income girls now taking the SAT (as compared to boys) are pulling the female averages down, it is ignoring the fact that girls at every income level score worse than boys with comparable family incomes (see Appendix H).

■ **Is It Possible to Create a Sex-Fair Test?: Construction of Sex-Biased and Sex-Equal Verbal Tests:** The existence of verbal SAT items that markedly favor one sex or the other on the SAT indicates that the 10 point "gender gap" suffered by girls nationally is manipulable by the content of the questions. Test-makers could easily construct a test on which one sex nationally scored as much as 50 points better than the other. On the June 1986 SAT, for example, if the 10 items that favored boys the most were deleted and replaced with items similar to the 10 items that most favored girls, girls nationally would outperform boys by about 4 points. This change would be accomplished solely with items that could pass through ETS's current screening process.

Since any difference between boys' and girls' means is dependent upon inclusion or exclusion of questions favoring one sex or the other, it is doubtful that the observed national 10 point difference can be considered "real" or that the test that created this difference can be considered "balanced." Instead, items could be included so that no difference in group

means for boys and girls would result. As ETS studies the performance of subgroups, items that particularly favor males, whites, and the affluent should be removed or balanced with items favoring females, people of color, and the working-class.

■ **Construction of Sex-Biased and Sex-Equal Math Tests:** As with the Verbal test, averages for males and females can be altered if existing math items favoring boys are replaced by items similar to current items that favored girls. Because boys outscored girls on most of the June 1986 SAT-Math items, a sex-equal math test cannot be constructed solely from existing questions. But, if the 10 most "pro-boy" items were replaced with items similar to the 10 most "pro-girl" items, boys nationally would outscore girls by about 29 points—thus eliminating more than a third of the existing gender gap.

For this study, the 4 questions favoring boys with the largest percentage differences were removed from both the Verbal and Math Sections of the November 1987 SAT and raw scores were recalculated to determine whether removing these questions would appreciably reduce the SAT gender gap. Although this made a difference on the Verbal Section, it did not affect the scores on the Math Section.

These findings—for both the June 1986 and November 1987 SAT—support the contention that ETS could construct a sex-equal Verbal test by including a relatively few more questions set in the context of experiences more familiar to females and eliminating a few of the questions that are most clearly set in a context familiar and comfortable to males. Since ETS tests all questions on the experimental sections of the test before using them, it should not be difficult to balance the Verbal Section. However, equalizing the Math Section appears to be more complex; extensive additional research may be needed to determine how this test can be made fairer to women and more predictive of their first year college performance.

■ **High School Achievement Tests—Are They Fair For Girls?:** Most high schools across the country administer standardized achievement tests to students at each grade level to measure their progress and to evaluate schools' performance. The 6 major tests are: the California Achievement Tests and Comprehensive Tests of Basic Skills published by CTB/McGraw-Hill; the Metropolitan Achievement Tests published by The Psychological Corporation; the Iowa Tests of Basic Skills published by The Riverside Publishing Company; and the Sequential Tests of Education Progress (STEP) and School and College Ability Tests published by ETS. According to CTB/McGraw-Hill product manager John Stewart, "very little bias was found on the California Achievement Test and those questions were balanced so that an equal number of items favored each sex." Questions also were analyzed by sex and race with a norming sample of African Americans and Hispanics in the same number or a greater percentage than their representation in the population in general.

■ **Girls' Score Averages Are Higher than Boys' on the Major Standardized Achievement Tests Used in High School:** Female/male performance differences on the California Achievement Test have also been studied extensively by Donald Ross Green, CTB/McGraw-Hill's Manager of Basic Research. In a representative sample of 110,000 students in grades K-12, he found that girls scored consistently higher than boys on most of the tests—in all ethnic groups examined (white, African American and Hispanic). Girls' higher performance resulted from better performance on almost all test items, rather than from a small group of items, while boys'

performance tended to be more variable than girls', for all ethnic groups studied.

■ **Longitudinal Studies—Cause for Concern:** However, the findings of two recent national longitudinal studies of high school performance show deficits in female performance similar to those in the SAT. These studies, conducted by the National Center for Education Statistics (NCES) and the federally funded National Assessment of Educational Progress (NAEP), raise questions about political intent; both studies used tests written by ETS and these findings are often cited by ETS researchers to justify the gender gap on the SAT. In the NCES study, high school senior girls had lost their lead over boys in reading and vocabulary; their reading performance was now similar to boys' and their vocabulary performance was lower.

All achievement tests *except* those for "High School and Beyond" (HSB) show girls outperforming boys in reading from age 9 onward, but as they get older the achievement gap narrows. The NAEP studies found that girls' reading proficiency at all three ages tested (9, 13, and 17) was declining in the 1980s, while boys made steady gains, narrowing the reading proficiency gap. This is particularly troubling, as reading is an area in which girls traditionally have received higher scores.

NAEP assessments of mathematics found few sex differences at ages 9 and 13, but males outperformed females at age 17, even when general course background was held constant. On HSB math tests, boys outperformed girls as sophomores and seniors, but girls earned higher average math grades, even in advanced math courses. In both NAEP's and HSB's writing assessments, girls clearly performed better than boys, with no changes in the size of the differences between the sexes over the years.

Other achievement test trends appear more ominous. In 1986, state-wide testing of high school juniors in Maine found large gender differences, with boys outperforming girls in math, science, and social studies. Girls outscored boys in reading, the humanities, writing, and writing mechanics. Again, researchers should question the purpose of achievement assessments that do not correlate with girls' superior classroom performance in math, science and social studies.

■ **The Narrowing of Cognitive Differences:** Sex stereotypical differences are currently being countered by other studies that show a narrowing of cognitive differences between the sexes. Yale Professor Alan Feingold found that gender differences had declined "precipitously" over the years on both the PSAT and the Differential Aptitude Test (DAT). The important exception was the "well-documented gender gap at the upper levels of performance on high school mathematics which has remained constant over the past 27 years."

Two important meta-analyses of tests by Janet Shibley Hyde and Marcia C. Linn have also found cognitive gender differences disappearing in verbal ability. Verbal differences were so small that they could "effectively be considered to be zero." The one outstanding exception was female performance on the SAT-Verbal, where the gender difference has been increasing. Their 1988 meta-analysis of gender differences in mathematics (not yet published), also found that math differences between the sexes were small. The largest differences occurred on questions that drew on advanced coursework in math and were similar to the gender differences in course enrollment for these subjects. Since most national assessment differences were declining, Linn and Hyde suggest that the "large, consistent gender differences found for the voluntary SAT-M sample are anomalous."

The evidence that achievement tests predict classroom grades equitably for both sexes is conflicting but these test results appear to be less damaging to girls' educational opportunities than the SAT, PSAT or ACT. It is not clear why girls find standardized achievement tests administered at high school grade levels less difficult but they seem to show that multiple choice tests are not *a priori* more difficult for females.

■ **Review of the Literature on Gender Bias in College Entrance Examinations:** Both the literature review and the comprehensive bibliography included in this report cite works that either contain direct references to the SAT, ACT, or Achievement Tests; refer to the issue of gender bias with regard to widely used basic skills tests administered to high school students; or focus on broader or related issues in ways that are immediately relevant to the study of gender bias in college entrance examinations.

***Chapter 3—Closing Doors: The Impact of Sex-Biased Tests on Women's Educational Opportunities*:**

■ **Sex Bias in National Merit Scholarship Awards**: Over $23 million in National Merit Scholarship awards, provided by 670 corporations, foundations, professional organizations, colleges and universities, are given annually to students with the highest scores on the Preliminary Scholastic Aptitude Test (PSAT). In 1987-88, women's average PSAT scores were 54 points lower than men's (13 points lower on the Verbal and 41 points lower on the Math); women therefore were only 36 percent of the National Merit Scholarship semifinalists while 60 percent of the semifinalists were men (some students' gender could not be determined by their names). In 1986-87, 34.7 percent of the semifinalists were women.

The semifinalist pool from which National Merit finalists and scholarship winners are chosen is based solely on the results of the PSAT administered to high school juniors each October. Students' PSAT scores must also be replicated by SAT scores in order for them to qualify as National Merit Finalists, so the bias on both these tests means that less scholarship money is awarded to girls. Talented young women also lose the prestige conferred on scholarship Semifinalists and Finalists that enhances college acceptance.

■ **Using SAT Scores to Award State Merit Scholarships: State-by-State Analysis:** Almost half (22) of the States offer merit scholarships to high school seniors who choose to attend colleges or universities in their home state. A state-by-state survey of the 1988 awards was conducted as part of this study to see whether girls were receiving a fair share. In States where SAT scores are used in combination with grades and class rank, or are not used at all, girls generally receive more scholarships than boys. In States where SAT or ACT scores are used exclusively, boys are more likely to receive scholarships.

New York awards the most state merit scholarship money of any State—$8.24 million annually. In 1988, the New York State Department of Education changed from using SAT scores only to a 50/50 formula of SAT (or ACT) scores and high school Grade Point Average to select scholarship winners. However, confusion in the reporting of grades resulted in girls receiving only 37 percent (compared to 28 percent in the preceding, SAT-only year) of the 1000 Empire State Scholarships of Excellence ($2,000 per year for 5 years) and 50 percent of the Regents Scholarships ($250 per year for 5 years), even though girls were 53 percent of the test takers.

When the State Department of Education decided to return to the exclusive use of the SAT in 1989, the Women's Rights Project of the

American Civil Liberties Union brought suit on behalf of the Girls Clubs of America, the New York chapter of the National Organization for Women, and 10 New York high school girls with grade point averages above 90. The suit charged that women receive unequal consideration because they tend to score an average of 60 points lower than men on the SAT while consistently earning higher grades in New York's high schools. Since the purpose of the Empire State and Regents scholarships is to reward outstanding high school performance, not to predict first year college grades—the avowed purpose of the SAT—this seemed an unfair criterion for determining scholarship winners.

Although the Education Department acknowledged that the SAT was not a perfect indicator of high school performance, it maintained that grades cannot be compared among schools because of grade inflation and because the collection process is too time consuming. Federal District Judge John M. Walker did not agree, ruling that the use of SAT scores as the sole basis for awarding merit scholarships is unequal treatment of girls; he enjoined the New York State Department of Education from awarding merit scholarships to high school students based solely on their SAT scores. Judge Walker found that this use of the SAT discriminates against girls "in violation of Title IX and the equal protection clause of the U.S. Constitution" (See Appendix I for the full text of the Opinion and Order).

■ **The Spin-off Effect:** Winners of State Merit Scholarships and National Merit Scholarships receive dozens of letters offering "no-need" scholarship awards, used by many colleges and universities to recruit high scoring students to attend their institutions. This spin-off effect is impossible to assess because it varies from student to student and state to state. However, it is important that parents and educators become aware of the interwoven nature of scholarship awards, in order to understand and appreciate the full extent of the financial and psychological damage inflicted by tests that do not predict classroom performance but do ensure access to important academic opportunities.

■ **Using SAT Scores to Choose "Gifted and Talented" Students: State-by-State Analysis:** Many states offer publicly funded academic enrichment programs during the summer to middle and high school students with high grades and high SAT, PSAT, or ACT scores. A State-by-State survey was conducted as part of this study to determine whether girls' educational opportunities at the middle and high school level were being limited by the use of these tests to select participants.

Seventeen States use SAT, PSAT or ACT scores as part of their admissions formula. However, these test scores generally are used as 20 to 30 percent of an evaluation portfolio that includes grades, essays, teacher recommendations, extra-curricular activities and demonstrated interest in the subject. Test scores therefore do not have an adverse affect on girls' participation in these summer programs; more girls than boys attend these programs, but involvement by both boys and girls of color is fairly limited. In fact, the evaluation process used by many States provided impressive alternatives to the exclusive or 50/50 use of college admission test scores.

■ **Private "Gifted and Talented" Programs—Exclusive Reliance on SAT Scores and Its Impact on Girls:** In contrast to these State programs, privately-funded summer programs for academically-talented 8th through 12th graders are far less open to girls. In the ten years since Johns Hopkins University began identifying "mathematically-precocious" children by administering the SAT to 7th graders, a number of similar talent search programs have been developed around the country. Academically-talented

students are usually identified as those who score 430 or over on the SAT-Verbal and 500 or over on the SAT-Math as 7th graders; the score cut-off goes up 20 or 30 points for each grade above 7th. These students are then invited to attend a summer camp offering accelerated courses in math, science and the arts at the university sponsoring the talent search.

Six Talent Search programs (based on the Johns Hopkins model and described in Chapter 3) were surveyed for this study, to assess the impact of girls' lower SAT score averages on their participation. It was not surprising that fewer girls participated in every program that used SAT scores for admission. One program—the ROGATE New Jersey Talent Search—used high school achievement tests instead of SAT or PSAT scores and had 2,018 females and 1,835 males participating in the 1988 summer program. Since more males than females participated in all the other programs, it would seem that the use of SAT scores is keeping girls out of privately-sponsored summer programs for "gifted" students. Although it was impossible to determine the exact number of programs now operating in the country, it appears that an increasing number of girls are affected by these talent searches.

■ **College Admissions—Are SAT Scores Essential?:** The SAT or ACT is required for admission to nearly all of the 1,500 four-year colleges and universities in the country. Many use strict cut-off scores, while others use test results in an admissions formula or require minimum SAT or ACT scores for admission to competitive departments or Honors programs. Nearly every college in the country publishes average SAT scores for its incoming first year class and parents and high school guidance counselors use them to assist students in college selection.

College admissions officers often use a mathematical formula that combines high school grades and SAT scores, weighting them in a way that predicts how well students are supposed to do in their first college year. If the same equation is used for both sexes, girls are predicted to do less well in college than they *actually* do (by one-fourth to a full standard deviation below their actual GPA), according to a 1973 study by the American College Testing Program.

Some young women in the June 1986 sample with A+ GPAs but lower SAT scores had self-selected themselves out of the elite college pool. They were not planning to apply to the most competitive colleges at the same rate as boys with similar grades. In fact, girls in all 4 GPA areas studied planned to go to slightly less prestigious colleges than boys with equivalent GPAs.

■ **Princeton University—A Case Study of Underprediction:** Even women who are accepted at the most competitive universities find their SAT scores underpredicting their college performance. In an unpublished senior thesis, Princeton University student Julie Lubetkin compared the grades, courses and SAT scores of the Princeton University Class of 1990, and found that the women's average SAT scores were slightly higher than the men's in the Verbal Section but considerably lower in the Math. Despite lower SAT scores, women's average first year GPAs were slightly higher than men's. In other words, SAT scores underpredicted the women's grades and overpredicted the men's grades, with the SAT-Math being the significant underpredictor.

■ **The Massachusetts Institute of Technology, Bates College, and Bowdoin College—New Admissions Policies To Counter the SAT's Underprediction for Women:** Some universities have taken action against the SAT's underprediction of women's academic performance. The

Massachusetts Institute of Technology's Admissions Office conducted a study of student performance and discovered that women with lower SAT-Math scores were achieving Grade Point Averages equal to or better than their male peers in their sophomore and senior years. According to Admissions Director Michael Behnke, this study "excluded the possibility that it is due to differences in course selection by men and women. Women also have a higher retention rate so it is not due to women dropping out at a higher rate." As a result, MIT has been admitting women with lower SAT scores than men.

Several other colleges have dropped the use of the SAT altogether, including Bates and Bowdoin in Maine, Middlebury College in Vermont and Union College in Schenectady, New York. Bates College found that applicants who chose not to submit SAT scores averaged 80 points lower on both the SAT Verbal and Math Sections than applicants who submitted their scores, but they did not differ significantly in first year GPA or academic standing. According to Wiliam A. Mason, Bowdoin's Director of Admissions, "in a climate where parents, guidance counselors and school boards all overemphasize the importance of test scores, we believe that our process is the fairest."

Chapter 4—Recommendations for Further Research and Development: The following are brief summaries of the Recommendations that conclude this report.

■ **Recommendations for Test Publishers:** Because ETS procedures proved unable to identify sex-biased items on the two SATs studied, different procedures are needed to reduce test bias:

1) ETS should eliminate from future SAT Verbal and Math tests those questions that show the largest gender, race, and class differences (see Chapter 2). Removing items from the test that have large response differences between the sexes, unless they are balanced by other items, is a first step towards achieving balance and fairness without compromising test integrity.

2) Since male and female mean scores on the verbal test are arbitrary and manipulable by the test-maker, the test-maker can manipulate them so that males and females score equally well, based on ability and knowledge; this would contribute to development of a sex-equal verbal test.

3) ETS and other test publishers should publicize the validity studies they now conduct on the relationship between SAT scores and first-year college grades and should make their findings available not only to other researchers but also to consumers.

4) ETS and other test publishers also should perform more research correlating performance on each SAT question with college grades.

5) ETS and other test publishers should allow test takers more time for each section of the test, to overcome the problems inherent in speeded tests, especially for women and students of color.

■ **Recommendations for Further Research:**

1) Conduct research on the predictive validity of the SAT and ACT for the college performance of women and men of color, including African American, Asian American, Hispanic, and Native American students of all socioeconomic levels.

2) Investigate the connections between sex and race bias in the classroom and bias in testing, to further assess the extent to which the SAT measures and therefore values the skills and knowledge that still differentiate upper middle class white males from others.

3) Conduct research on the impact of coaching on women and girls, students of color, and low income students.

4) Conduct further research on test anxiety to investigate why girls are more anxious, so that steps can be taken to decrease their anxiety.

5) Conduct further research to explain one of this study's most surprising and distressing findings: that the largest sex differences in SAT score averages—much larger than the national averages for the test as a whole—occurred between boys and girls with the highest high school grades (A+ to A), while the smallest gender gap occurred at the lowest GPA level.

6) Conduct research that would contribute to development of useful, predictive, and fair alternatives to standardized testing to evaluate students' achievements and predict their future performance.

CHAPTER ONE

Defining Sex Bias in Standardized Testing

Introduction

Standardized tests are multiple choice examinations administered to large sample populations to determine average, above average and below average performance for certain types of skills. They are widely used as achievement tests in elementary and secondary schools to evaluate academic progress and to identify students in need of compensatory education. They are also used as aptitude tests in the college admissions process to predict a prospective student's first year grades.

Sex bias may be inherent IN the test itself or may be a result of the way in which the test is used; bias can be expressed in four ways:

- in *test content*, when many more men than women are referred to or depicted and women are shown in lower status or stereotyped roles (facial bias);
- in *test context*, when questions are set in experiences more familiar to one sex than the other; women and girls, for example, tend to prefer questions with aesthetic-philosophical and human relations content while boys/men prefer questions dealing with science and the world of practical affairs (Strassberg-Rosenberg and Donlon, 1975);
- in *test validity*, when women's academic abilities are underpredicted by test scores while men's are overpredicted; and,
- in *test use*, when women's access to educational opportunities is diminished or restricted by an institution's reliance on a test that underpredicts their abilities.

There has been convincing evidence for the past 15 years that standardized tests used for college admissions are biased against women in all four areas. Although test content has become fairer, the underprediction of women's academic abilities has gradually grown worse, decreasing their opportunities in both admissions and scholarships. However, public concern has evolved slowly.

Initial research on *test content* was conducted by Professor Carol Kehr Tittle, currently Director of the Doctoral Program in Educational Psychology at the City University of New York. In 1973, Tittle found that many educational tests referred to males much more frequently than to females, showed men in higher status positions and depicted both sexes in stereotyped roles; women, for example, nearly always were shown at home or in the pursuit of hobbies, as if the professions were closed to them (Tittle, 1974). As Tittle stated, even if these depictions of a male-oriented world did not have a negative effect on girls' test scores, they are offensive in their perpetuation of cultural bias against females and should be eliminated. Several studies have shown that women are more likely to succeed on a question when the

people depicted are either female or "neutral" in sex; yet men continue to outnumber women in items on many tests (Selkow, 1984), including the Scholastic Aptitude Test (SAT) which is widely used for college admissions decisions (Rosser, 1987).

The form of sex bias that has the greatest negative impact on women's educational opportunities is the underprediction of their first year college performance by both of the major college admissions tests—the SAT, published by Educational Testing Service (ETS) and taken annually by 1.5 million students (52 percent of whom are female), and the American College Testing Program's ACT Assessment, taken annually by nearly one million students (54 percent of whom are female). The major purpose of these tests is to predict first year college grades, but studies show that women receive higher grades in all subjects in both high school and college classes from their first year onward. Yet, they receive lower scores on both the SAT (College Entrance Examination Board, 1988) and the ACT (Gamache and Novick, 1985). They also receive lower average scores on the Preliminary Scholastic Aptitude Test/National Merit Qualifying Test (PSAT/NMQT), published by ETS and taken annually by approximately 1.1 million high school juniors, 54 percent of whom are female. Although the PSAT is defined as a "practice test" for the SAT, the National Merit Scholarship Corporation awards over $23 million in student scholarships each year to the students with the highest scores on this test, making it extremely important as a "gateway" to college for many students.

The Impact of Sex-Biased Tests on Women's Educational Opportunities

Reliance on these biased tests has an adverse impact on young women's educational opportunities in three important ways. By underpredicting their academic performance, the tests affect women's chances to gain entrance to nearly 1500 four-year colleges and universities that require SAT scores or use SAT cut-off scores for admission (Rosser, 1987). Unfairly low test scores also become a self-fulfilling prophecy for many girls and young women; lower scores inspire lower expectations and encourage women to apply to less competitive colleges and universities than their grades otherwise would warrant. In fact, a 1987 Carnegie Foundation report found that 62 percent of the students questioned had lowered their college expectations after receiving their SAT or ACT scores (Boyer, 1987).

Lower test scores also exclude girls from academic enrichment programs and accelerated courses open only to students with the "top" test scores. A number of summer programs are offered by state universities, private colleges and universities, and well-known preparatory schools; only seventh

through eleventh graders who score 430 or higher on the Verbal Section of the SAT and 500 or higher on the Math Section are eligible. SAT or ACT scores are also used as the admissions criteria for state-sponsored summer programs for academically-talented high school students.

Reliance on biased tests has a severe economic impact on women, who lose millions of dollars in merit scholarship awards, despite their higher grades. Merit scholarships are awarded annually by 22 states as well as by hundreds of corporations, foundations, professional organizations, unions and government agencies, based on SAT, ACT or PSAT scores. Although most of these organizations refuse to provide a gender or racial breakdown of scholarship recipients, the National Merit Scholarship Corporation—which offers the most prestigious awards for academic excellence—selects its semifinalists *solely* on the basis of PSAT scores. In 1987-88, women's PSAT scores averaged 44 points lower than men's and their qualifying scores (the verbal score doubled with the math score added) were 67 points lower—thus leaving women eligible for only 36 percent of the approximately 6,000 scholarships.

The continuing result is a significant dollar loss for women in later life as they get less prestigious jobs, earn less money and have fewer leadership opportunities. Of course, the life-long loss of self-confidence cannot be measured in financial terms.

The Underprediction of Women's Academic Performance by the SAT

The SAT is composed of two sections, Verbal and Math, each scored on a 200-800 point scale; the maximum possible combined score is 1600. In 1988, women's average SAT scores were 56 points lower than men's—13 points on the Verbal Section, an area where women excelled until 1972 when men began to outscore them—and 43 points on the Math Section, where women have always scored lower then men (Dwyer, 1976a). However, the College Board's own Validity Studies show that women's average first year college grades are as good or better—in all subjects—than are those of their male peers, who have higher SAT scores (Clark and Grandy, 1984).

This would suggest that the SAT is not fulfilling its primary purpose—to predict first year college performance—for women. Indeed, if the SAT *were* predictive, these young women would either earn lower first year college grade point averages than they actually do or they would receive higher average test scores than men—perhaps 10 or 20 points higher rather than 56 points lower. Since 52 percent of the test taking population is female, this test is underpredicting grades for approximately 780,000 young women every year. It is significant but little known that when young men were

receiving lower verbal test scores, even without higher grades, the SAT-Verbal test was rewritten, according to the College Board, to improve the gender balance (Donlon and Angoff, 1971).

The College Board, which administers the SAT, reported in 1988 that women who took the test had a far higher mean Grade Point Average (GPA) than the men who took the test (CEEB, 1988). As the chart below indicates, of students with the highest grades (A+), 53 percent were women, compared to 47 percent men. Women were 58 percent of test takers with A averages, 57 percent of students with A- averages, and 54 percent of B students. In contrast, men were the majority of test takers with C averages (56 percent) and D averages (64 percent).

While SAT scores for both men and women declined from 1973 to 1982, high school grade point average and class rank have been consistently higher for women than for men throughout these years (Clark and Grandy, 1984).

Male/Female Grade Point Averages

Grade Point Average	Percent of Females	Percent of Males
A+	53	47
A	58	42
A-	57	43
B	54	46
C	44	56
D or less	36	64

(Source: CEEB, 1988)

According to the College Board, the male score advantage on both sections of the SAT cannot be explained by cognitive differences. A recent College Board report states that "the research literature finds no difference between men and women in performance on cognitive skills, or finds a slight advantage for females on verbal skills and a slight advantage for males on mathematical and spatial skills. Male-female biological differences do not appear to explain the observed difference in cognitive functioning; experiences, stereotypes, and expectations no doubt play a role, but it has been difficult to identify specific ways in which they may account for differences in academic performance. In addition, the measures we use may contribute to the differences we observe" (Clark and Grandy, 1984). Although women score less than a point lower than men on the Graduate Record Examination (GRE), they receive higher verbal scores on nearly all other standardized aptitude and achievement tests.

The gender gap also cannot be attributed to large variations in academic preparation. The College Board reports that in 1988, 88 percent of the

women had taken four years of English compared to 86 percent of the men; 59 percent of the female test takers had completed four years of math compared to 68 percent of the males; 84 percent of the women had completed three years of social sciences, compared to 82 percent of the men; 72 percent of women had three years of natural science, compared to 79 percent of men; and 88 percent of women had taken two years of foreign language, compared to 82 percent of the men (CEEB, 1988). Looking at the percentage of male and female students who have taken algebra and geometry, reportedly all the math needed for the SAT, the figures are even closer. In 1987, 97 percent of both sexes had taken algebra; and 93 percent of the males and 92 percent of the females had taken geometry (CEEB, 1988).

The SAT's underprediction for women has not been a secret. ETS researchers Clark and Grandy (1984) state that "the underprediction of women's first year college grades has been reported consistently in the research literature" (p. 21). The 1988 College Board Report, *Taking the SAT, 1988-89,* warns against using the SAT alone to evaluate students: "SAT scores are intended to supplement the secondary school record and other information about the student in assessing readiness for college-level work" (p. 4).

A similar pattern of sex bias can be found on the PSAT/NMQT, which is also constructed with a Verbal and a Math Section. Each section is scored on a scale of 20 to 80; ETS claims that an approximation of future SAT scores can be obtained by multiplying scores by 10. In 1987-88, girls averaged 41 points lower on the Math and 13 points lower on the Verbal than boys. To qualify as a National Merit semifinalist, verbal scores are doubled and the math score is added, in order to "give girls a better chance." However, as girls' verbal scores decline, doubling the verbal score is not overcoming the large gender gap in math scores. To win a National Merit Scholarship, test takers must replicate their PSAT scores with their SAT scores, which also works against girls.

Women of color are doubly penalized by the SAT. They all score lower than the men in their racial/ethnic group, according to the latest College Board Report (CEEB, 1988). All men of color, in turn, receive lower combined average scores than white men. For example, African American women averaged 32 points lower than African American men in 1988 and 241 points lower than white men; African-American men averaged 209 points lower than white men.

Approximately 6,000 "nonstandard" SATs are administered to students with disabilities each year. Although the literature review conducted for this study did not find any studies that compared male and female differences, limited research has been done on the testing of students with visual, hearing, and physical impairments and students with learning disabilities. One ETS study found that visually impaired students and those with physical disabilities achieved average scores on the SAT, as compared to non-disabled students. In contrast, students with learning disabilities and hearing impairments did not perform as well as the general test-taking population. Time extensions were given to test takers with disabilities and tests were written in braille for visually impaired students. Not surprisingly, there tended to be a lower correlation between high school grades and SAT scores for students with disabilities than there is for the SAT test-taking population in general.

Underprediction by the ACT Assessment

Women's college performance is also being underpredicted by the ACT Assessment, the other major college entrance examination, which is taken by nearly a million students in the Midwest, Southwest and South. Fifty four percent of the test takers are female; college grades thus are being underpredicted for nearly 54,000 students. The ACT is considered an achievement rather than an aptitude test, surveying acquired knowledge in four subject areas: English Usage, Mathematics Usage, Social Studies and Natural Science. Each section of the test is scored on a scale that ranges from 1 to 36.

Men receive higher average scores than women in all subject areas except English Usage, while women continue to earn higher grades in these same subjects. Like the SAT, the ACT is also not accurately predicting young women's first year college grades. In 1987-88, the average ACT Composite Score for men was 19.9 compared to 18.6 for women (ACT, 1987). Gamache and Novick (1985) found that all ACT subject scores and the ACT Composite score (the average of the combined subject scores) "consistently underpredicted" women's two-year cumulative college Grade Point Average, even when partially controlling for different courses taken.

The ACT is also having an adverse impact on male and female students of color, who receive lower ACT composite scores than do white males. As the chart below indicates, women within each ethnic group receive lower scores than men.

1987-88 ACT Composite Scores

White males	21.0
Asian American males	20.8
White females	19.6
Asian American females	19.4
Puerto Rican males	18.1
Puerto Rican females	16.6
Native American males	16.6
Mexican American males	16.3
Native American females	15.1
Mexican American females	14.8
African American males	14.1
African American females	13.4

(Source: *College Student Profiles: Norms for the ACT Assessment*, 1987)

Although the ACT plays a major role in college entrance for many students, it has not been as intensely studied as the SAT because it is not preferred for entrance by the most elite colleges and universities and does not determine the winners of National Merit Scholarships. In January 1989, the American College Testing Program announced that the ACT Assessment has been redesigned to emphasize a wider range of mathematical knowledge, more abstract reading skills and a new testing of scientific concepts. Since the first redesigned ACT will not be administered until October 1989, no item analysis of this test can be included in this study.

CHAPTER TWO

The SAT Gender Gap: Identifying the Causes

Gender Bias in SAT Items: An Initial Assessment[1]

Men have always received higher scores than women on the SAT. Years ago, their higher math scores were partly offset by women's higher verbal scores (by approximately 5 points) (CEEB, 1987). But women lost their verbal lead in 1972, due to gradual changes in the test content that added questions referring to science, business, and "practical affairs" and eliminated questions with human relations, arts, and humanities content (Dwyer, 1976a). ETS changed the test to create "a better balance for the scores between the sexes" (Donlon and Angoff, 1971, pp. 25-26). As a result, by 1986 the verbal gender gap favored men by 11 points. Dwyer noted that a change in test specifications required more male-oriented items on verbal tests, where women traditionally excel, but the reverse (more female-oriented items on math tests, where men traditionally excel) had not been required; she called this "nonconscious sexism" (1976b).

To investigate this male advantage, an item analysis of one form of the June 1986 SAT was conducted, to determine whether specific questions were creating or widening the score gap between the sexes, to investigate other factors that might contribute to sex differences in SAT scores, and to determine how SAT scores influenced students' future academic plans.

This study sought to determine which test items, if any, showed marked gender-related biases favoring girls or boys; to investigate item-to-scale (point-biserial) correlations of sex-biased items, in order to study methods of test construction that might reduce sex bias; to investigate relationships among SAT scores, high school grade point averages (GPAs), and sex, to see if women's lower SAT scores were accompanied by correspondingly lower school performance; to investigate other factors, such as socioeconomic status, test anxiety, and high school subject preference, that might help explain why women do worse than men on the SAT but not in high school or college; and to investigate the effects of SAT scores on students' college choices and self-perceived abilities, by sex.

Methods

In March, 1987, 1,112 students in Princeton Review coaching classes took a form of the June 1986 SAT, during the second session of their coaching class, under conditions as similar as possible to those in ETS test centers. As

the final section of the exam, 1,028 students answered an additional 25-item questionnaire (Appendix B), which asked them to indicate their high school grade point averages (GPA), favorite high-school subjects, perceived ability in English and math, test anxiety, and family background.[2]

■ **Sample:** Because not every student answered every item, the sample size ranged from 1,112 on some SAT items to about 1,010 on some questionnaire items.[3] Students came from the five boroughs of New York City, from selective public high schools such as Bronx Science and Stuyvesant, nonselective public schools, parochial schools, and private schools such as Dalton. They were fairly closely balanced between the sexes; 55.6 percent were girls and 44.4 percent were boys, while nationally, SAT takers are 52 percent female. Most students (75.3 percent) were white, but the sample included 13.2 percent Asian Americans, 5.2 percent African Americans, 2.4 percent Hispanics, and 3.9 percent who checked "other" or left the column blank. Almost all (97.8 percent) were in the 11th grade; 0.7 percent were sophomores, and 1.5 percent were seniors.

Students' high school preparations were rather strong. In self-reported high school grade point average, 57 percent reported averages from B+ to A+. Nor were these grades earned in so-called "easy" courses. In math, including their current year's classes, 86 percent had taken three years, one course per year; 11.7 percent had taken more. English preparation was also strong; 92 percent of the girls and 91 percent of the boys had taken three years of English in their three years of high school and 7 percent had taken more. In natural science, 73 percent of all students had taken three years; among the rest, boys were more likely to have taken an additional year while girls were more likely to have taken less.

The students came mainly from upper-middle class backgrounds and reported that 81 percent of their fathers and 52 percent of their mothers had professional careers (doctors, executives, engineers, teachers, for example) and 72 percent of their fathers and 63 percent of their mothers were college graduates. Sixty percent of the sample attended public school, 9.5 percent parochial, and 27 percent prep school, with little difference between the sexes, except that 4 percent more males were attending parochial schools.

Because this sample came from one metropolitan area and selected themselves by paying for an expensive coaching course, they cannot be seen as random or representative of the national population. However, within this group, valid internal comparisons—boys versus girls, anxious versus not anxious—can be made and it is likely that the processes operating within this sample can be generalized to others. The uniformity of this sample regarding socioeconomic status is especially valuable as it allows an exploration of differences by sex that cannot be attributed to low incomes or lesser educational preparation.

Results: Sex Differences

Each of the SAT's 85 verbal items is worth about 7 score points and each of the 60 math items is worth about 9.5 score points.[4] On the verbal SAT nationally, men now outscore women by about 10 points, or 1.4 items; this sample was about equal. On the math scale, men outscore women nationally by about 47 points, or 5 items; among the students in this sample, males outscored females by 3.5 items or about 35 scale points.

TABLE 1

SAT Averages by Sex

Group	Verbal Raw	Verbal Scale	Math Raw	Math Scale	Total Scale
Female, National		425		453	878
Male, National		435		500	935
Female, Sample	44.8	489	33.6	536	1025
Male, Sample	45.0	490	37.1	571	1061

The 17 Questions with Major Sex Differences

Girls and boys scored within a few percentage points on most verbal and math items, reflecting the fact that wide areas of experience, skills and subcultural terms are shared by young people of both sexes, and that most SAT questions tap those areas. However, 7 items on the Verbal and 10 on the Math Sections of the SAT showed considerable (more than 10 percent) differences in the percentage of each sex that answered them correctly; (for the full text of these questions see Appendix A). Among the 85 verbal items, 22 additional items favored one sex or the other by more than 5 percent, a cut off point suggested by Green (1987). Table 2 below lists the verbal items with approximately 10 percent or greater differences. Those favoring girls are indicated by a + sign.

TABLE 2

The 7 SAT Verbal Items That Favored One Sex by Approximately 10 Percent or More

Section, Item No., Description	Female %–Male %
1 No. 1, "setback," opposite "improvement"	-10.7
1 No. 5, "sheen," opposite "dull finish"	+18.3
1 No. 23, author's tone, science passage	-11.8
1 No. 44, "mercenary is to soldier"	-15.7
4 No. 21, "pendant is to jewelry"	+ 9.6
4 No. 24, "love is to requite"	+14.5
4 No. 31, "betrayal" (in human relations item)	+10.2

In a society in which sex stereotypes still have an impact, it is not surprising that words referring to relationships ("requite"), jewelry ("pendant"), and fabric ("sheen") favor girls; conversely, "mercenary" relating to "soldier" is a male-loaded term in a society that drafts only men for military service. Previous studies (Coffman, 1961; Strassberg-Rosenberg and Donlon, 1975; Dwyer, 1979) have found that item content produces important sex differences in performance. In fact, recent public concern with item bias and wording has led ETS to create a Sensitivity Review Process (ETS, 1987) for all items. However, Table 2 indicates that the Sensitivity Review Process does not eliminate items that favor one group (see below for a discussion of problems with the ETS approach).

Among math items, 10 differences of greater than 10 percent appeared, all favoring men; these differences are marked (-) in Table 3.

[redacted]

TABLE 3

The 10 SAT Math Items That Favored One Sex by More Than 10 Percent

Section, Item No., Description	Female %-Male %
2 No. 8, "liters per hour"	-10.3
2 No. 15, "chore 994th boy will have at boys camp"	-12.3
2 No. 16, "number of boy with chore at boys camp"	-15.6
2 No. 19, "parallelogram ratios"	-12.2
2 No. 20, "1/6 as decimal, sum of digits"	-10.7
2 No. 21, "basketball team won/loss record"	-27.0
2 No. 22, "<(a-b)<"	-11.0
2 No. 25, "n as odd integer"	-10.8
5 No. 17, "length of right triangle"	-10.7
5 No. 25, "inequalities with x^2, -x"	-10.6

Three math items—numbers 15, 16, and 21—were specifically about boys' enterprises, suggesting that verbal bias adversely affects girls' performance on math items. Earlier studies have shown that when math content is made relevant to female experience, males do not outperform females on math problems (Milton, 1958; Bem and Bem, 1970; Graf and Riddell, 1972; Donlon, 1973; McCarthy, 1975). Items on which boys markedly outperformed girls ranged in difficulty from easy to hard, implying that the level of mathematics involved did not cause the difference in performance by sex. Among the 60 math items, 16 additional items favored one sex or the other by more than 5 percent.

The Students Most Affected by Sex-Biased Items

Researchers conducting this study suspected that middle-range scorers would be most affected by sex-biased items. High scorers might be more likely to be certain of the right answer, while low scorers might not know anything about the right answer, so they might not even guess at it. Middle scorers might know just enough to guess, but their "subliminal" knowledge would be more easily misled by sex-biased items. Table 4 divides the sample

into four groups by overall verbal scores (low = 200-480, low-middle = 481-530, high-middle = 531-580, and high = 581-800), and again by math scores (200-520, 521-580, 581-650, and 651-800).[5] For all items listed in the previous tables, Table 4 displays the mean absolute differences (percent of males answering the items correctly minus percent of females, for male-biased items; the reverse for female-biased items). As hypothesized, middle scorers were affected most, though the differences were small.

TABLE 4

Mean Differences by Sex in Percentage Correct on Sex-Biased Items, Among Low, Middle, and High SAT Scorers

	SAT Score Range				
	Low	Low-Mid	High-Mid	High	All
Among Verbal Items:	14.3	14.1	16.5	7.9	13.0
Among Math Items:	4.6	6.0	7.9	6.9	-12.0

Do Certain Types of Questions Favor One Sex?

ETS divides verbal items into four types: antonyms, reading comprehension questions, sentence completions, and analogies. Contrary to studies that found that women (Strassberg-Rosenberg and Donlon, 1975) and African Americans (Schmitt and Dorans, 1987) do better on reading comprehension items and worse on analogies (Donlon, 1973; Stricker, 1982), this study found that women and men performed about the same on all item types. Women did better at antonyms and worse at reading comprehension, but the differences were slight. There were no important differences by difficulty of item on the verbal test.

For the purposes of this study, math items were classified into four types: computation, geometry, algebra, and problem solving. Prior research has been equivocal as to which sex does relatively better on which types of questions. Donlon (1973) found that women performed relatively better in algebra than geometry, while Milton (1957) and Graf and Riddell (1972) found that problem solving favored men. Becker (1983) found SAT algebra items more difficult for junior high girls than boys, but no sex differences in geometry and computation. McPeek and Wild (1987) found women performing better on algebra than geometry on the Graduate Record Examination (GRE).

This study found nothing to substantiate consistent sex differences; girls scored closer to boys on computation, but the difference was slight and they performed no better on algebra compared to other math areas. Nor did this study find important differences by difficulty of item; boys outscored girls on all but 8 items, although the differences were predictably smallest on the easiest items. Table 5 indicates that the type of question differentiated between males and females much less than did item content, as shown in Tables 2 and 3.

TABLE 5

Scores by Sex on Different Types of Items

Type of Questions	Average Percentage Correct		
	Female	Male	Female % - Male %
Antonyms	62.2	60.9	1.3
Reading Comprehension	46.5	47.8	-1.3
Sentence Completion	71.0	71.5	- .5
Analogies	66.0	65.6	.4
10 Easy Verbal Items	87.5	88.5	-1.0
10 Medium Verbal Items	59.6	57.9	1.7
10 Difficult Verbal Items	25.4	25.4	0.0
26 Algebra Items	62.5	67.8	-5.3
14 Geometry Items	54.1	58.8	-4.7
14 Computation Items	71.5	74.9	-3.4
6 Word Problems	60.6	65.8	-5.2
10 Easy Math Items	85.9	86.4	-0.5
10 Medium Math Items	55.7	63.2	-7.5
10 Difficult Math Items	19.2	28.0	-8.8

Do SAT Sex Differences Correlate with Performance Differences?

Since SAT scores could not be correlated with first year college grades for this sample, a surrogate was used for this study: high school GPA. Researchers have consistently found that high school GPA is the best single predictor of college GPA, and although its *r* is only about .48, that is higher than *r*'s for the SAT or most other predictors (ACT, 1973; CEEB, 1987).

In this sample, SAT scores correlated only moderately with high school GPA.[6] Girls in this sample are performing considerably better in high school than their relative SAT scores would suggest. Although they received lower scores than boys on both parts of the SAT, Table 6 shows that they earn higher grades. Thus, this SAT is underpredicting girls' high school GPAs.

Another way of showing this underprediction is to try to use SAT scores to predict high school GPAs, by sex. Table 7 shows that the SAT "predicts" high school GPA well, within each sex, but with marked female/male differences. Within almost every SAT score category, looking across the top data row, a higher percentage of girls earn A to A+ grades than boys. For example, 41.7 percent of girls with top Math SATs earn A to A+ grades, while only 31.4 percent of boys do. This trend even continues for B+ to A- grades, which is surprising since there are so many girls in the top row that fewer remain at the lower grade point levels.

TABLE 6

Percentage Reporting Various High School GPAs, by Sex

GPA	Percentage of Girls	Percentage of Boys
A to A+	18.7	15.3
B+ to A-	43.0	36.8
B- to B	30.8	39.8
C+ or Lower	6.0	7.4

TABLE 7

Percentage Reporting Various High School GPAs, by SAT Score Range and Sex

	Percentage of Girls With Verbal SATs				Percentage of Boys With Verbal SATs			
Percent with GPA of:	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
A to A+	4.2	14.3	24.1	34.8	4.4	12.9	17.7	27.9
B+ to A-	37.6	41.4	43.6	50.4	24.8	34.7	40.6	49.2
B- to B	43.0	36.1	29.3	12.8	56.9	43.6	37.5	19.7
C+ or Lower	12.7	6.8	2.3	0.7	13.8	7.9	2.1	3.3

	Percentage of Girls With Math SATs				Percentage of Boys With Math SATs			
Percent with GPA of:	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
A to A+	3.1	12.0	25.3	41.7	2.6	4.7	14.5	31.4
B+ to A-	30.6	47.3	51.3	43.5	14.1	28.0	48.1	45.7
B- to B	49.4	36.0	19.5	12.0	60.3	59.8	31.3	21.4
C+ or Lower	14.4	4.7	1.9	0.9	23.0	6.5	5.3	0.7

The study also compared grades in high school English courses with Verbal SAT scores and grades in high school math courses with Math SAT scores. Again, controlling for SAT scores, more girls earned A to A+ grades, compared to boys, in both English and math. These findings agree with CEEB validity studies cited by Clark and Grandy (1984) that show women receiving college grades equal to or better than men's in math, science, and the humanities. The Massachusetts Institute of Technology (MIT) has also found that women with lower SAT Math scores earn college grades equal to those of men; MIT therefore has changed its admissions policies accordingly to limit the influence of SAT scores (Behnke, 1987).

Student Factors That May Cause Sex Differences in SAT Scores

■ **Test Anxiety:** Researchers have suggested that test anxiety may create different performance by sex on the SAT. Indeed, young women in this study reported considerably more test anxiety, as Table 8 shows.

TABLE 8

"How Do You Feel About the SAT?" by Sex

Level of Anxiety	Females	Males
"extremely anxious"	27.8%	10.8%
"moderately anxious"	38.5	37.7
"somewhat anxious"	24.9	34.2
"not anxious at all"	8.8	17.3

There were 2 1/2 times as many "extremely anxious" girls as boys. Girls' anxiety may constitute a rational response to their history of lower SAT performance, compared to their high school grades. However, in this sample, test anxiety did not correlate closely with poor test performance, particularly among boys. Among girls, the least anxious group scored considerably worse than others. "Extremely anxious" girls scored lower on the Math SAT than "somewhat" anxious girls, but anxiety levels had no effect on verbal scores.[7]

High school GPA also had no systematic relationship with test anxiety, nor did students' own rating of their verbal and math skills. However, the more anxious the test takers, the more likely they were to believe that tests underrate abilities. Socioeconomic status also influenced anxiety; students whose fathers were professionals were less anxious than those whose fathers were not professionals. Mother's occupation made no difference to sons, but 33.7 percent of daughters of women who do not work outside the home were "extremely anxious" about the SAT, compared to 23.1 percent of daughters of mothers with professional careers. Parents' education had

mixed impact, but generally, children of more educated parents were less anxious. Anxiety correlated moderately with plans to attend "super-elite" colleges.

■ **Time Pressure:** Graf and Riddell (1972) found that on math problems perceived to be more difficult, girls proceeded more slowly than boys, but others have found no appreciable sex differences in test-taking speed (Donlon, 1977; Wild, Durso and Rubin, 1982). To determine whether either sex was more affected by time pressure, we examined performance on the last 10 items on the last two tests, Section 4 (Verbal) and Section 5 (Math), and found no important differences by sex. Girls did slightly better than boys on the final verbal items; boys did better on the final math items, just as they did on earlier math items. Almost identical percentages of boys and girls left the last 5 items blank on the Verbal test; on the Math test, 5.5 percent more girls left the items blank, but slightly more girls than boys left earlier math items blank as well.

■ **Liking Mathematics Helps Math SAT Scores:** To explain the large gender gap in math scores, researchers have suggested that prior sex stereotyped socialization influences boys to like math more than girls and to take more math courses in high school. More than 36 percent of the boys in this sample chose math as their favorite subject or chose science first and math second, compared to 22.4 percent of the girls. Another 13.2 percent of boys and 11.6 percent of girls chose math as their second favorite subject.

Liking math raised scores on the Math SAT for both sexes, as Table 9 shows, but the male/female gap remained, though it narrowed somewhat. Among students who reported that they liked math, for instance, males held a 2.6 point advantage, while in total scores, *all* males had been 3.5 points ahead. Of course, liking math may also partly be a *result* of good scores on prior "standardized" tests and may also correlate with course-taking.

TABLE 9

Math SAT Items Correct by Math as Favorite Subject, by Sex

Mean Number of Items Correct	Math First	Math Second	Math Not Chosen
Among girls:	40.9	38.5	36.0
Among boys:	43.5	41.4	37.7

Interestingly, students of both sexes who chose math as their favorite subject earned lower scores on the Verbal SAT. This would be understandable if one assumes that students who like math do not like English; it is not understandable if, on the other hand, one believes that math is a difficult subject and that students who like math might be more studious, hence better in all subjects. We also found several items on which math-

likers did *much* worse (more than 10 percent) than math-dislikers; only one favored math-likers by 10 percent. These findings suggest the need for further study.

■ **Taking Mathematics:** ETS notes that the Math SAT does not utilize math beyond algebra and geometry, so that students who had taken more advanced math courses should not be advantaged simply by that fact. Table 10 shows that most of the students in this sample—regardless of sex—had taken one year of math per year in school. Only 23 students had omitted a year or more of math. Although 18 of these were girls, the percentage of all girls taking less than the typical three years of math was only 2.9 percent; 15.7 percent of the boys took more than 3 years of math, compared to 8.6 percent of the girls, which is fairly similar to national studies (Ramist and Arbeiter, 1986). However, extra math did not affect SAT performance substantially, probably because higher math is not required for SAT Math questions. Table 10 primarily indicates that most of the students in this sample have had one year of math in each year of high school, regardless of gender.

TABLE 10

Math Preparation by Sex

	Percent of Students		
Years of Math in High School	All	Girls	Boys
4 or more	11.7	8.6	15.7
3 (one/year)	86.1	88.5	82.8
2 or less	2.2	2.9	1.5

(Not adjusted for the 2.5 percent non-juniors)

Table 11 shows the relationship of years of math to SAT scores. Unlike "likes math," "takes math" does not adversely affect Verbal SAT scores, while it does correlate with higher Math SAT scores. Controlling for years of math taken slightly narrows the gender gap in Math SAT scores. The largest group in the sample, students in the "3 Years" column, show a 2-point gap, a bit less than the 3.5-point gap in the entire sample. Girls taking less math than average exhibit a 5-point deficiency compared to boys. Like a National Assessment of Educational Progress (NAEP) study (Welsh, Anderson, and Harris, 1982), the young men in this study still did somewhat better after the effects of differential preparation were removed.

Interestingly, taking math correlates with better performance on some verbal items but not on others. Moreover, on some items, taking math correlated with better performance for girls more than for boys, while on others, the reverse was true.

TABLE 11

Math SAT Items Correct by Amount of Math Taken, by Sex

Mean Number of Items Correct	Years of Math Taken		
	4 Years	3 Years	<3 Years
Among girls:			
Verbal SAT	52.6	51.0	49.4
Math SAT	42.1	37.6	33.9
Among boys:			
Verbal SAT	51.6	50.7	51.9
Math SAT	43.8	39.5	39.0

Socioeconomic Factors That May Cause Sex Differences in SAT Scores

■ **Parental Education:** Like other research on SAT performance, this study found that fathers' and mothers' education levels and occupations had immense impact on their children's scores.[8] Daughters and sons of more educated fathers (who had more than a bachelors degree) averaged about 8.5 more Verbal SAT items correct compared to children of less educated fathers (who had not attended college). Daughters of educated fathers did better on all but 6 of the 85 items, and more than 10 percent better on 39 items, than daughters of less educated fathers. Sons of educated fathers did better on all but 5 of the 85 items, compared to sons of less educated fathers, and more than 10 percent better on 48 items. In math, daughters of more educated fathers averaged 5.3 more items correct than daughters of less educated fathers; sons varied by 10.3 items. Daughters of educated fathers did better on all but 3 of the 60 math items, and more than 10 percent better on 20. Sons of educated fathers did better on all but 9 of the 60, and on those 9 did about the same, while they did more than 10 percent better on 24 items.

Mothers' education correlated with better performance for boys even more than girls on the Verbal test; sons of more educated mothers got 11.5 more verbal items correct, compared to sons of less educated mothers, while daughters varied by 8.6 items. Daughters of more educated mothers did better on all but 2 Verbal SAT items compared to daughters of less educated mothers, and they did more than 10 percent better on 39. Sons of more educated mothers did better on *every* Verbal SAT item, and they did more than 10 percent better on 55 questions. On math items, children of more educated mothers got 5.7 more math items correct than children of less educated mothers. Daughters of more educated mothers did better on all but one Math SAT item compared to those of less educated mothers, and they

did more than 10 percent better on 28 of the 60 items. Sons of more educated mothers did better on all but 3 Math SAT items compared to those of less educated mothers; on 27 items, the difference was greater than 10 percent.

■ **Parental Occupation:** Fathers' occupations correlated with SAT scores, confirming ETS's consistent reporting of high positive correlations between parental income and SAT scores. Fathers' occupations made about twice as much difference for boys as for girls. Mothers' occupations made the same kind of difference as fathers', as Table 12 shows; children of professionals had higher scores than children of mothers with "other" occupations. Mothers' occupations hold additional interest owing to the category, "works in home," which reflects a "traditional" role for women. And it correlated with a big difference in scores, especially among girls. Table 12 indicates that *not* working outside the home had about the same effect as holding "other" occupations (real estate, social worker, sales clerk, waitress, for instance).

TABLE 12

Mean Number of SAT Items Correct Correlated to Mother's Occupation, by Sex

	Among Students With Mother's Occupations:			
	Professional	Other	Works in Home	Diff.(Col.1-3)
Girls, Verbal	54.5	48.2	48.8	5.7
Boys, Verbal	55.3	50.1	48.7	6.6
Girls, Math	38.1	36.1	36.7	1.4
Boys, Math	42.8	38.7	39.9	2.9

Girls whose mothers worked only in the home perceived their English ability to be lower than girls whose mothers had professional careers, and scored lower on the Verbal SAT, although their high school grades in English were equal. This suggests a link between SAT scores and girls' perceptions of their mothers' status in society, which may translate to girls' self esteem. It also may be that mothers who work at home have lower self esteem which they may pass on to their children, a possibility suggested by the research of Jacobs and Eccles (1985) on math ability.

Among boys, mother's occupation did not correlate with perceived English ability. Regarding perceived math ability, the picture reversed: mother's occupation made little difference to daughters, but did correlate positively to sons' perceptions and to sons' math grades. These findings suggest the pervasive influence of social class on students' scores; this is all the more striking in view of the constricted social class range among the families of students in this sample.

Of course, class influences on SATs have been pointed out many times before; class also underlies some (although not all) of the gap between African American and white scores. On some items, students from higher income families scored more than 10 percent above others; indeed, on several items, they scored more than 20 percent better. On other items, socioeconomic status made little difference. This may suggest that some items are "classist" in the same way that some have proven to be sexist and some to be racist. The importance of improving ETS's item-selection process to promote gender fairness would hold even more strongly regarding class and race.

Effects of the SAT on Students

Students in this sample had a good self-image regarding their own abilities. In "reading and writing ability," girls and boys were almost identically positive; a majority (57.3 percent) placed themselves in the top 10 percent of their peers, while only 1.7 percent believed that they were in the bottom half. In math ability, girls were less sure: 38 percent, compared to 56 percent of boys, claimed to be in the top 10 percent.

The students also showed healthy self-images or serious criticism of "standardized" tests; in response to the question—"Do you feel your past test scores on standardized tests (PSAT, etc.) are accurate?"—81.3 percent claimed their "ability is higher than the tests indicate." There were no important sex differences.

SAT Differences, High School GPA Differences, and Perceived Ability, by Sex

Standardized "aptitude" tests can adversely affect students' self-image, as many students with lower test scores may reasonably believe that they have low "verbal aptitude" or "math aptitude," since ETS uses "aptitude" to title its tests. Almost all of the students in this sample had taken ETS tests previously, and their scores on the SAT can be taken as a surrogate for prior scores. Test feedback was compared to teachers' evaluations of student performance (high school grades), to see which had the greater impact on students' own reports of their verbal and math abilities. Tables 13 and 14 show ratios of girls' scores to boys'. When girls and boys were equal, the ratio is 1; if girls scored better, the ratio is >1; if girls scored worse, the ratio is <1.

Girls earned better grades in English than boys, but had comparable scores on the Verbal SAT; they ranked their English abilities only a little higher than boys, in line with the SAT results. In math, girls did about as well in school, but worse on the Math SAT; again, they estimated their math ability in line with the test results, not the classroom results. Thus, although girls and boys earned almost identical grades in math, only 38 percent of girls put themselves in the top 10 percent in math ability, compared to 56 percent of boys. These findings confirm Clark and Grandy's findings (1984), that students' overall perceptions are closer to test feedback than to grade feedback, which is beneficial for boys' self image but damaging to girls'.

Students compare SAT scores at least as avidly as grades. Moreover, students can provide *reasons* for poor grades—not doing the homework, not studying. But for poor SAT scores, students can only supply excuses: "I don't do well on 'standardized' tests," "I don't care about it anyway," "I had a bad day," or the assumption that "I'm not good in math." It is likely that some

TABLE 13

Ratio of Girl/Boy Ranking on English High School GPA, Verbal SAT Scores, and Perceived Verbal Abilities

Item	Girl Result Divided by Boy Result
% A+ on English HS GPA	1.58
% A- through A+ on English HS GPA	1.20
% in Highest Group on Verbal SAT	.95
% in Highest Two Groups on Verbal SAT	1.01
% Estimating Their Verbal Ability in Top 5%	1.11
% Estimating Their Verbal Ability in Top 10%	1.05

TABLE 14

Ratio of Girl/Boy Ranking on Math High School GPA, Math SAT Scores, and Perceived Math Abilities

Item	Girl Result Divided by Boy Result
% A+ on Math HS GPA	.85
% A- through A+ on Math HS GPA	.96
% in Highest Group on Math SAT	.62
% in Highest Two Groups on Math SAT	.77
% Estimating Their Math Ability in Top 5%	.52
% Estimating Their Math Ability in Top 10%	.69

girls internalize the SAT's underprediction of their academic performance as an assessment of their "aptitude."

Self-perception and test performance are probably *inter*-dependent. Table 15 sheds light on this point; it shows the same strong relationship between SAT score and self-perceived ability that previous tables have displayed. In "reading and writing ability," self-perception and SAT scores are similar for both sexes—49.6 percent of girls who scored well on the Verbal SAT rank themselves in the top 5 percent, for instance, compared to only 40.2 percent of high-scoring boys, but the difference is made up in the next category, the top 10 percent.

TABLE 15

Percent of Students Who Place Themselves In the Listed Percentile Groups in Self-Perceived Abilities, as Affected by SAT Scores

	Among girls				Among boys			
	Low	Low-Med	Med-High	High	Low	Low-Med	Med-High	High
				VSAT Groupings:				
Self-perceived reading and writing ability								
top 5%	12.1	15.8	22.6	49.6	8.0	18.8	24.0	40.2
top 10%	29.1	37.6	39.1	31.2	22.6	36.6	35.4	41.8
top 25%	29.7	24.1	25.6	14.9	38.7	29.7	28.1	13.9
top 50%	21.2	14.3	3.8	0.7	25.5	12.9	9.4	1.6
bottom 50%	4.2	3.0	0.0	0.0	2.9	1.0	1.0	0.8
				MSAT Groupings:				
Self-perceived math ability								
top 5%	2.5	7.3	16.2	38.9	5.1	13.1	21.4	57.1
top 10%	9.4	22.7	32.5	36.1	9.0	22.4	41.2	30.7
top 25%	32.5	36.0	35.1	16.7	28.2	34.6	23.7	10.7
top 50%	28.8	25.3	7.1	2.8	42.3	22.4	13.0	1.4
bottom 50%	20.6	6.7	0.6	0.0	11.5	3.7	0.0	0.0

In perceived math abilities, however, the sexes behave differently. Girls have a lower perception of their abilities even when they do well on the Math SAT. Among high scorers, for instance, 57 percent of boys put themselves into the top 5 percent in math ability, while only 39 percent of girls did so. Conversely, among low scorers, 20.6 percent of girls put themselves in the lower half in math ability, while only 11.5 percent of boys did so. In other words, when the test tells them they are good at math, girls are less likely than boys to believe it, and they are more likely to believe that low scores reflect their ability.

On the other hand, among girls in the *lowest* scoring group on the MSAT, 89 percent say their "ability is higher than the tests indicate," while 81 percent of the low-scoring boys agree. Thus, girls do not simply internalize low MSAT scores.

Do SAT Scores Influence Future Aspirations?

Students displayed high college aspirations with more than 95 percent of both young men and women planning to attend "super-elite," "very strong," or "strong" four-year institutions. High school grades largely determined whether students planned to attend "super-elite" rather than "very strong" colleges. Sex made some independent difference, as Table 16 shows, with only 52 percent of A to A+ girls planning to attend "super-elite" colleges, compared to 66 percent of A to A+ boys. Among B+ to A- students, the difference is even greater; twice as many men (36.3 percent) as women (18.7

percent) plan to attend "super-elite" schools. The fact that SAT scores matter most to applicants to competitive "super-elite" institutions implies that the lower SAT scores received by girls with very high grades, compared to boys with very high grades, might have contributed to girls' lower aspirations.

TABLE 16

Students Who Plan to Attend Different Types of Colleges, by High School GPA

	Students with High School GPAs of:			
Type of College	A to A+	B+ to A-	B- to B	C+ or lower
Percent of female students choosing:				
Super-Elite	52.3	18.7	7.4	11.8
Very Strong	30.8	48.4	34.1	11.8
Strong	14.0	30.1	55.7	52.9
Percent of male students choosing:				
Super-Elite	65.7	36.3	7.7	5.9
Very Strong	22.9	40.5	31.9	20.6
Strong	8.6	20.2	56.6	50.0

However, when we looked directly at the influence of SAT scores on college choices, we found that they did not account for the sex differences in choice of "super-elite" schools. Within each SAT score category, boys were more likely to attend "super-elite" colleges than girls. Among high MSAT girls, for example, 45 percent plan to attend "super-elite" colleges; among high MSAT boys, 51 percent planned to do so; the difference, 6 percent, is exactly the same as between *all* girls (see Table 17); therefore, we cannot lay the difference at the doorstep of the SAT.

Summary of Major Findings

This study confirmed the underprediction that other researchers have noted: girls received lower scores than boys on the SAT, yet they had higher high school grades than boys in both English and math. Further, the study found significant item bias; 17 items were considerably (more than 10 percent) easier for one sex, suggesting that ETS's review process does not work as effectively as it should. Specific item content made the greatest difference, rather than the type of item (an analogy), the subject matter (geometry), or the level of difficulty.

The study also found that girls' poorer performance was not linked to test anxiety or time pressure, which often are postulated as reasons for women's poorer scores. While boys liked math somewhat better and took slightly more math, this only explained part of their Math SAT lead over girls; in addition, liking math adversely affected Verbal SAT scores to some extent. Controlling for social class still produced a score gap favoring boys; thus,

TABLE 17

Students Who Plan to Attend Different Types of Colleges, by SAT Scores

	Students with Verbal SATs				
Type of College	Low	Low-Mid	High-Mid	High	All
Percent of female students choosing:					
Super-Elite	5.5	14.3	19.5	47.5	21.2
Very Strong	29.1	44.4	42.1	39.0	38.1
Strong	58.2	37.6	34.6	11.3	36.4
Percent of male students choosing:					
Super-Elite	11.7	18.8	26.0	51.6	27.0
Very Strong	27.7	28.7	46.9	31.1	32.9
Strong	51.1	50.5	25.0	13.9	35.5

	Students with Math SATs				
	Low	Low-Mid	High-Mid	High	All
Percent of female students choosing:					
Super-Elite	5.6	14.7	26.6	45.4	21.2
Very Strong	25.6	39.3	51.3	36.1	38.1
Strong	61.3	42.0	18.2	17.6	36.4
Percent of male students choosing:					
Super-Elite	6.4	5.6	30.5	51.4	27.0
Very Strong	15.4	40.2	37.4	32.9	32.9
Strong	62.8	51.4	30.5	12.9	35.5

social class did not explain the gender gap. Independently, socioeconomic status had a high correlation with SAT scores—children of parents with higher status jobs and more education scored better.

Finally, when estimating their math and English abilities, both men and women perceived their abilities to be more in line with their test scores than with their grades. Unfortunately, this meant that girls believed themselves to be less able than their grades would indicate, and less able than boys. And girls were less likely to aspire to "super-elite" colleges. Further, sex differences in these two areas also persisted when SAT scores were controlled for, with men ranking their abilities moderately higher than women and aspiring to "super-elite" colleges at a moderately higher rate, suggesting the complexity of the operation of sex bias in education.

Implications for Test-Makers

■ **Reviewers Cannot Reliably Detect Biased Content:** ETS used three procedures to evaluate items during the construction of this SAT: (1) using deltas to assess the general difficulty of each item to assemble tests containing the desired number of easy, medium, and difficult questions: (2) reviewing item content; and, (3) calculating item-to-scale (biserial) correlation coefficients.

ETS's descriptions of its item review process (Donlon, 1984; ETS, 1987; cf Donlon and Angoff, 1971) do not make clear the details of the process as applied to a given test. Apparently, proposed items are reviewed to see that they do not offend any ethnic groups or either sex. Perhaps they are also reviewed to see that they do not obviously favor the subculture and vocabulary of any "subgroup of English speakers" (ETS, 1987). The results of this study's preliminary item analysis, however, challenged the effectiveness of this face validity check. Verbal items with sex stereotyped content, ("pendant" and "mercenary," for instance), were left on this exam and proved to favor one sex or the other by considerable margins.

To assess the effectiveness of ETS's procedure, Loewen replicated it, judging each VSAT item for male or female bias, simply on the basis of subject matter, before looking at any results. Loewen predicted that girls would do better on 7 items, boys on 3. Results proved his predictions correct on 9 items and wrong on only 1. It is surprising that the ETS review process could not identify and eliminate culturally-loaded items that were noted by a single untrained observer. This is especially problematic as an ETS researcher made similar predictions and achieved similar results more than a quarter century ago, before ETS's review process was in place (Coffman, 1961). Although his judging of items was more effective than ETS's, Loewen missed several on which one sex scored more than 10 percent better than the other. The content of one item, "sheen" opposite to "dull finish," obviously drew upon the subculture and vocabulary of girls. But other items on which one sex showed a peculiar advantage were not so obviously biased in content, particularly on the math test.

Our knowledge as to differences in vocabulary and cognitive styles among different racial groups and between boys and girls is modest; hence, even after our results flagged an item as favoring one sex or the other, it was not always possible to explain why. Therefore, it is doubtful that sex bias (or racial or class bias) can be predicted consistently on the basis of item content. However, ETS could examine the performance of men and women of each racial/ethnic and class group on each item after they have appeared on the experimental section of the SAT, to determine which questions actually create the largest differences.

On the Math Sections, similar problems occurred—3 of the questions on which boys showed the greatest advantage dealt with boys' camp and basketball team statistics; yet Rosser's item analysis of the November, 1987 SAT (described below) did not find that math questions set in the context of female experience advantaged girls. Nonetheless, Loewen's experience suggests that face validity review is not an effective procedure to detect and remove biased items.

■ **Item-to-Scale Correlations Cannot Detect Bias:** After items have been judged fair, or at least inoffensive, ETS includes them on experimental sections of the SAT and computes item-to-scale *r*'s. Such correlations have no mitigating effect on sex or racial bias. Indeed, to the degree that the test as a whole favors affluent, white, or male subcultures, using *r* to screen items will maintain or increase bias on sex, class, or racial lines.

An example can clarify this point. Imagine a verbal SAT item that tapped working-class culture, such as item 3, "Spline is to miter as straw is to mud," from the "Loewen Low-IQ Test" (Loewen, 1979). It involves difficult reasoning and might help predict which students from working-class culture were most capable of that reasoning, but it would never get past the biserial *r* hurdle, because upper- and middle-class students would get it wrong, while some working-class students would get it right. Since SAT scores are

strongly class-related, "spline" would not correlate well with overall scores. Hence, no item favoring working-class culture is likely to be included on any SAT. Indeed, we found that point-biserial *r*'s for "classist" items were higher than for class-fair items on this test.

The situation is similar regarding sex and the Math SAT. Because girls score worse than boys, any item on which girls excelled would be unlikely to have a robust biserial *r*, so ETS would drop it. Indeed, we note that the 5 most "pro-boy" items on the Math SAT show *r*'s averaging .45, while for the 10 items on which girls approximately equalled boys, average $r = .30$.[9] Indeed, the *r* test probably acts to increase sex bias on the Math SAT. On the Verbal SAT, using the biserial *r* to qualify an item has no systematic effect on sex bias, because boys and girls are roughly equal in numbers and performance. Thus "pro-boy" and "pro-girl" verbal items can pass this hurdle and be included. A "pro-girl" math item would probably not make it onto the test, nor would a "pro-minority" item. Recommendations for new procedures to avoid item bias, which were included in the original version of this study, have been incorporated into the recommendations that conclude this report.

Gender Bias on the November 1987 SAT—An Item Analysis

This item analysis is based on the responses of 100,000 college-bound high school seniors to one form of the November 1987 Scholastic Aptitude Test (SAT), contained on a College Board data tape compiled by the Educational Testing Service (ETS); this sample represented nearly all the students who took one of the four forms of the test administered at that time. According to ETS, the test forms are distributed throughout the country in such a way that one form of the test is taken by students from all ethnic and income groups in every geographic area. Therefore, this is the best random sampling of the student population that ETS makes available to the public.

The results of this item analysis represent a substantial new body of data to explain the causes of the gender gap in SAT scores. This research is among the first by an independent researcher, not affiliated with ETS, that—using ETS data—attempts to determine whether specific questions create or contribute to the score gap, whether the SAT correlates with current academic performance for both sexes, and whether other factors might be causing sex differences. This study met three primary objectives:

- To identify and analyze the questions that showed large differences in performance between men and women in general and in every racial/ethnic group. Although women in every racial/ethnic group receive lower average SAT scores than the men in their group (College Board, 1988), almost no research has been done in this area.

■ To see how well SAT scores correlated with current high school performance for both sexes, particularly for boys and girls who reported the highest grade point averages (A+ to A)—to determine whether the SAT predicted current performance as accurately for the high achieving women as it did for high achieving men. These are the students who rely most heavily on the SAT to gain entrance into the "elite" colleges and universities.

■ To investigate other factors—such as risk taking, time pressure and socioeconomic status—that might explain why girls receive lower average SAT scores but higher average grades than boys.

Do Some SAT Questions Show Large Performance Differences by Sex and Race?

■ **Sex and Race Differences in SAT Score Averages:** On the Verbal SAT nationally, boys now outscore girls by about 10 points. On the Math SAT, boys outscore girls nationally by approximately 50 points. The national score averages for all SATs administered to the class of 1988—which includes every SAT taken by these college-bound seniors prior to April 1988—are close to the score averages of the students on the data tape (College Board press release, Tuesday, September 20, 1988). Although 52 percent of the members of the class of 1988 were female and 48 percent were male, this sample was 55 percent female and 45 percent male. Their average scores are compared in Table 18.

TABLE 18

SAT Averages by Sex

National SAT Averages for the Class of 1988

	Women (590,299)	Men (544,065)	Difference
Verbal	422	435	-13
Math	455	498	-43
		TOTAL DIFFERENCE	-56

SAT Averages by Sex for November 1987 100,000 Sample

	Women (54,606)	Men (45,391)	Difference
Verbal	431	445	-14
Math	462	506	-44
		TOTAL DIFFERENCE	-58

Women received lower average scores than men on both sections of the SAT; they averaged 14 points lower than males on the Verbal Section and 44 points lower on the Math. This is consistent with the pattern of SAT Score Averages reported by the College Board since 1981 (College Board press release, Tuesday, September 20, 1988). Further, women in every ethnic group received lower average scores than the men in their ethnic group, as shown in Table 19. The largest score gap occurs between Hispanic women

TABLE 19

Score Averages by Sex and Race, With Female Difference Within Racial/Ethnic Group [November 1987 Sample]

Group (Number)	Verbal	Math	Combined Score	Female Difference
White Males (33,620)	457	517	974	
White Females (40,846)	444	473	917	57
Asian American Males (2,694)	416	543	959	
Asian American Females (2,724)	409	503	912	47
Native American Males (429)	410	471	881	
Native American Females (601)	396	418	814	67
Hispanic Males (1,791)	398	457	855	
Hispanic Females (2,373)	380	406	786	69
African American Males (2,829)	381	414	795	
African American Females (4,441)	371	388	759	35
Other Males* (4,028)	429	485	914	
Other Females* (3,621)	417	447	864	50

*Did not indicate race.

and men (69 points) and the smallest occurs between African American women and men (35 points). Earlier College Board research (Ramist and Arbeiter, 1986) had also found this consistent score gap.

Table 20 ranks each racial and gender group by combined SAT scores to show which groups received the highest and lowest scores on the test. Although white males received the highest average scores (974) and African American females the lowest (759), Asian American males averaged 26 points higher than white males on the SAT-Math. Asian American females averaged only 14 points lower than white males on the SAT-Math, in contrast to white females, who averaged 43 points lower. This finding raises interesting questions about potential differences in the preparation of girls of different racial/ethnic backgrounds in mathematics.

TABLE 20

Ranking of Each Racial/Ethnic Group by Combined Scores, from Highest to Lowest [November 1987 Sample]

White Males	-	974
Asian American Males	-	959
White Females	-	917
Asian American Females	-	912
Native American Males	-	881
Hispanic Males	-	855
Native American Females	-	814
African American Males	-	795
Hispanic Females	-	786
African American Females	-	759

The average verbal and math scores for males and females in each racial group were combined to arrive at average SAT scores by race, shown in Table 21.

The Questions with Major Gender Differences

Of the 145 questions on the test, 23 displayed substantial differences in the number of women and men who answered them correctly or a large difference in the proportion (ratio) of females to males who answered them correctly. A closer analysis was conducted of all questions with an approximately 10 percent or greater difference between females and males in the percentage of correct answers. The study also looked at the ratios of females to males answering each question correctly; ratios of .699 and lower were chosen as an indicator of bias, as women performed less than 70 percent as well as men in their answers.

In the Verbal Section, girls scored considerably lower than boys on 4 questions and higher on 2 questions, as shown in Table 22; for the full text of these questions see Appendix D.

TABLE 21

Average SAT Scores for Each Racial Group, Highest to Lowest, for November, 1987 Sample

Percent of Total Students Taking Test	Verbal	Math	Combined Score
White students 74.47%	450	494	944
Asian American students 5.42%	414	523	937
Native American students 1.03%	402	440	842
Hispanic students 4.16%	387	428	815
African American students 7.27%	375	398	773
Other students 7.65%	423	467	890

TABLE 22

6 SAT VERBAL Items Favoring One Sex by Approximately 10 Percent

Section, Item Number—Description	Female % - Male %	Ratio
1, No. 2—Opposite of IRK is SOOTHE	+ 8	1.111*
1, No. 37—Reading Comprehension passage about the orbit of a comet	-10	0.737*
4, No. 2—Opposite of STAMINA is LACK OF ENDURANCE	-12	0.867*
4, No. 4—Opposite of SHEEPISH is CONFIDENT	+ 9	1.145*
4, No. 25—Sentence Completion about sports	-25	0.390
4, No. 41—Analogy—Dividends:Stockholders	-15	0.717*

*Ratios bigger than cut-off of .699

As our preliminary item analysis (reported above) also found, a larger percentage of women than men chose the correct answers for questions referring to relationships ("irk" and "sheepish") and a larger percentage of men chose the correct answers for questions referring to physical science, sports, and the stock market. ETS researchers Wendler and Carlton (1987) have found that girls perform better on questions that are general and abstract or set in a context—a characteristic of humanities questions. Boys perform better on questions that are specific and concrete—characteristics of questions about science and practical affairs. Again, ETS's attempt to "balance" Verbal content with equal references to areas that interest each sex has not been attained. As an example, in Section 4, Question 25, not only is the percentage difference between the sexes unusually large but the ratio of females to males answering the question correctly is unusually low. The question is shown below:

> 25. Although the undefeated visitors——triumphed over their underdog opponents, the game was hardly the—— sportswriters had predicted.
>
> (A) fortunately ..... upset
> (B) unexpectedly ..... classic
> (C) finally ..... rout
> (D) easily ..... stalemate
> (E) utterly ..... mismatch

Two-thirds more boys than girls answered this question correctly—an extreme difference that makes this an especially inappropriate question for the predictive purpose of this test. Questions such as this one, set in the context of sports journalism, have no relation to academic abilities; the fact that these stereotyped "boy" topics are unfair to girls on the SAT further suggests that they should be eliminated from the test.

Among the 60 Math questions, 17 exhibited large (10 percent or more) percentage or ratio differences between the sexes, *all* favoring men (Table 23). In fact, men outscored women on every math question on this test, despite their lower average math grades. The differences were smallest on the easiest questions (at the beginning of each section) and largest on the most difficult items (at the end of each section).

The pattern in math word problems is worth noting. As Table 23 shows, young women found 6 of the 10 word problems on the test (numbers 2/8, 2/17, 2/21, 5/10, 5/30, and 5/31) considerably more difficult than did their male peers, supporting research over the years that has found that math word problems prove more difficult for females (Graf and Riddell, 1972; Donlon, 1973; Chipman, 1988). However, it is important to note that earlier research also has found that males do not outperform females on math word problems if the problems are set in content familiar to females (Milton, 1958; Bem and Bem, 1970; Graf and Riddell, 1972; Donlon, 1973; McCarthy, 1975). The item analysis on this SAT, however, does not support this finding. Three of the problems are about food or cooking and one is about a female making pottery plates, all of which seem to be content relevant to traditional female experience. Yet girls also performed worse than boys on these questions, as they did on others; this result suggests that women may find word problems, regardless of content, more difficult than other types of math. This raises questions, therefore, about both math curriculum and pedagogy as well as about the test itself.

Research conducted during the past two decades has shown that even though teachers of both sexes believe that they are treating boys and girls

TABLE 23

17 SAT MATH Items Favoring One Sex by More Than 10 Percent or a Large Ratio

Section, Item Number—Description	Female %-Male %	Ratio
2, No. 7—"If 2/3 of n is 4, then 1/2 of n is"	-13	0.827*
2, No. 8—"Pat made a total of 48 pottery plates"	-12	0.848*
2, No. 12—"If x = 80 and y = 30, what is the value of k?"	-11	0.847*
2, No. 17—"If the least possible multiple of the recipe"	-10	0.697
2, No. 18—"which of the following points on the square"	-10	0.643
2, No. 21—"how many plants will there be in 1989?"	- 9	0.609
2, No. 23—"Lines Q 1 and Q 2 are not parallel"	- 9	0.690
2, No. 24—"letters opposite each other are reciprocals"	- 5	0.688
2, No. 25—"what is the solution for $x^2 + x + c$"	- 6	0.600
5, No. 6—"which of the following pairs of numbers"	-16	0.784*
5, No. 10—"The number of gallons of gas if tank is 75% full"	-10	0.877*
5, No. 18—"The average (arithmetic mean) of x, y, and z"	-12	0.826*
5, No. 29—"What is the value of x in triangle division?"	-13	0.783*
5, No. 30—"If the price of mints was raised from 5 cents"	-10	0.863*
5, No. 31—"If a rectangular cake is cut into x equal rectangles"	-11	0.761*
5, No. 33—"how many different-sized circles"	-11	0.607
5, No. 35—"If s equals 1/2% of t, what % of s is t?"	- 4	0.429

*Ratios bigger than cut-off of .699

similarly, there are subtle differences in their expectations for and behavior towards each sex (Sadker and Sadker, 1985; deNys and Wolfe, 1985). Textbooks also perpetuate this "hidden curriculum" of sexism by rarely portraying women and people of color in non-stereotyped ways. This and other analyses of the SAT suggest that efforts must be made to ensure that standardized tests do not reflect this "hidden curriculum" in ways that perpetuate harm to women and girls.

■ **Do Certain Types of Questions Favor One Sex?:** To determine whether preferences for different subjects (algebra over geometry or reading comprehension over analogies, for instance) were creating the score gap, men's and women's average percentage of correct answers in each skill area of the test were compared (see Table 24). In addition, the average percentage of correct answers for the "easy," "medium" and "difficult" questions were compared to ascertain whether significant sex differences existed. "Easy" questions were defined as the ones for which 70 to 100 percent of the students chose the correct answer; "medium" questions were the ones that 40 to 69 percent of test takers answered correctly; and "difficult" questions were answered correctly by 39 percent or fewer of the students. Girls performed slightly better than boys on the easy verbal items and somewhat worse on the difficult items but the difference was not large. Large gender differences did appear, however, in comparing the "easy," "medium" and "difficult" questions in the Math Sections.

TABLE 24

Easy, Medium, Difficult Items

Type of Questions	Average Percent Correct		
	Female	Male	Female % - Male%
10 Easy Verbal Questions	86.20	85.00	+1.20
10 Medium Verbal Questions	51.10	52.00	-0.90
10 Difficult Verbal Questions	17.30	21.90	-4.60
10 Easy Math Questions	83.80	88.30	-4.50
10 Medium Math Questions	51.70	58.60	-6.90
10 Difficult Math Questions	16.00	23.00	-7.00

ETS divides verbal questions into four types: analogies, antonyms, sentence completions, and reading comprehension passages. Earlier studies have found that women perform better on reading comprehension questions and antonyms and worse on analogies and sentence completion questions (Donlon, 1973; Strassberg-Rosenberg and Donlon, 1975; Wendler

and Carlton, 1987); but this item analysis found that girls performed slightly worse on all item types. The differences on all item types were small, although they were somewhat larger for analogies (-2.55 percent) and sentence completions (-3.13 percent) than for antonyms (-1.16 percent) and reading comprehension questions (-1.56 percent); these results are presented in Table 25.

The math questions are classified into four types: arithmetic, algebra, geometry and other (graphs, set theory, series, and probability). Past research has found that girls perform less well in geometry than in algebra or arithmetic (Donlon, 1973; Strassberg-Rosenberg and Donlon, 1975; and McPeek and Wild, 1987) and the findings of this item analysis confirm this. Geometry questions had the largest sex difference of all the item types on the test; the male average percentage correct for these questions was 8.8 percent higher than the females'. Arithmetic questions showed the smallest math difference between the sexes, with the male average percentage correct 4.86 percent higher than the females'. Earlier research found that SAT arithmetic items favored girls (Strassberg-Rosenberg and Donlon, 1975; Fennema and Sherman, 1977) raising the question of what is happening now to cause this change.

TABLE 25

Scores by Sex on Different Types of Items

Type of Questions	Average Percent Correct		
	Female	Male	Female % - Male%
25 Antonyms	50.84	52.00	-1.16
25 Reading Comprehension	40.92	42.48	-1.56
15 Sentence Completion	63.47	66.60	-3.13
20 Analogies	48.75	51.30	-2.55
27 Algebra Questions	48.56	55.26	-6.70
15 Geometry Questions	39.93	48.73	-8.80
14 Arithmetic Questions	66.07	70.93	-4.86
4 Other Math Questions	63.00	69.00	-6.00

■ **The Math Score Gap:** The mathematical score gap between the sexes is not a recent development. It has been present on the SAT at least since 1967, when the College Board first published national data on college-bound seniors; apparently it has always existed. According to Carol A. Dwyer, ETS Senior Development Director for Test Development, "efforts have not been made to 'balance' the SAT quantitative sections, even though sex differences

have favored males by a great number of points since the first administrations of the test" (Dwyer, 1976a). Despite the fact that women earn consistently higher grades in math classes, in both high school and college, the SAT-Math gender gap has ranged from 41 to 52 points for the past 21 years.

A recent study by Gross and Sharp of more than 4,000 high school students in Montgomery County, Maryland public schools, found that girls who took the same advanced math courses as boys—calculus, pre-calculus and advanced algebra—even in the same classrooms and with the same teachers, earned higher grades but lower Math SAT scores; in fact, the girls' SAT scores were 37 to 47 points lower than the boys' scores (Gross, 1988). In a study of Rutgers University's class of 1985 first year students, which included more than 1,000 women, Ellen Kanarek found that the women had higher average grade point averages than the men in science and math; their GPAs in the humanities were substantially higher than the men's (Kanarek, 1988).

But, in a recent meta-analysis of gender differences in mathematics, Marcia C. Linn and Janet S. Hyde found that gender differences are declining on other national assessments. The largest differences between the sexes were found in questions that drew on advanced coursework and were "similar in magnitude to the gender differences in enrollment in these courses." Given these declines, they say that "the large, consistent gender differences found for the voluntary SAT-M sample are anomalous" (Linn and Hyde, 1988).

The fact that female performance on the Math Section of the SAT has always been worse than males', despite women's higher math grades, and that "balance" has not been attempted (Dwyer, 1976a) or achieved, raises important questions about the intent of the test publishers. Test questions are written to meet the publisher's content specifications; what decisions have ETS test developers made to justify specifications that produce questions that do not predict female performance?

How can consumers (college admissions officers) address these problems? To give only one example, the Massachusetts Institute of Technology (MIT) has decided to deal with the math score gap by admitting women with lower SAT-Math scores than their male peers. The admissions office examined the predictive validity of SAT-Math scores, comparing them to college grades; they found that while women had significantly lower scores on the SAT-Math, there were no significant differences between male and female grade point averages. As a result, MIT has decided not to restrict admissions to students who score over 750 on the Math SAT; 60 percent of MIT's first year class in 1986 scored below 750, with 8 percent scoring 500 or below (Behnke, 1987).

■ **Do women choose different wrong answers than men?:** It has been suggested that men's and women's different cognitive styles affect their success on the SAT and in different intellectual endeavors. Therefore, the wrong answers chosen by females and males were examined for the questions with the largest percentage and ratio differences, to determine whether cognitive style could be creating the score differences. However, females and males chose the same wrong answers (distractors) in about the same proportion for each question, suggesting that both sexes use similar thought processes to answer the test questions (see Appendix E).

Questions Showing Large Sex Differences Within Each Racial/Ethnic Group

African American women exhibit the smallest gender gap and Hispanic women the largest, when compared to men within their own racial/ethnic group. This study sought to determine which questions were creating the problems and whether there was a discernible pattern. Table 26 shows all the questions that had a 10 percent or greater difference in percentage of correct answers or large ratio differences between males and females in each racial group—whites, African Americans, Asian Americans, Hispanics, Native Americans and Others (students who did not indicate racial background).

Only one verbal question made a large difference for women in every racial group—the sentence completion question (item 25 in Section 4) which relates to sports journalism. The four verbal questions that created a gender gap on the test overall also created a gap for white women, the largest group of test takers. Only two verbal questions had a larger than 10 percent difference for Asian American males and females: the sentence completion question (item 25 in Section 4) relating to sports and the analogy (item 41 in Section 4) "dividends:stockholders as royalties:writers."

Three verbal questions showed large differences for African American women compared to African American men: "the opposite of 'mobile'" (item 1 in Section 1); "the opposite of 'stamina'" (item 2 in Section 4); and the sports sentence completion item (25 in Section 4).

Seven verbal questions made a difference for Hispanic women, compared to Hispanic men: "the opposite of 'mobile'" (item 1 in Section 1); "the opposite of 'mottled'" (item 14 in Section 1); "All are correct statements about Comet Brooks except:" (item 37 in Section 1); "the opposite of 'stamina'" (item 2 in Section 4): "the seizing of Cherokee lands" (item 23 in Section 4); the sports sentence completion item (item 25 in Section 4); and, "dividends: stockholders as royalties:writers" (item 41 in Section 4).

Seven verbal questions also differentiated between Native American women as compared to Native American men: "the opposite of 'mottled'" (item 14 in Section 1); "Comet Brooks is like Halley's" (item 40 in Section 1); "the opposite of 'stamina'" (item 2 in Section 4); "the opposite of 'affirmation'" (item 6 in Section 4); "the opposite of 'inter'" (item 15 in Section 4); the sports sentence completion item (25 in Section 4); and, "dividends:stockholders as royalties:writers" (item 41 in Section 4).

A total of 38 math questions created a gender gap for women of color. The mathematics gender gap was smallest for African American women; although they scored lower than any other ethnic/gender group on the test, only six math questions showed differences of more than 10 percent or large ratio differences compared to African American men. Native American women had the largest math gender gap, with 24 questions that had substantial percentage or ratio differences. Hispanic women followed with 22 questions, white women with 18 and Asian American women with 16.

Fourteen of the 17 math questions which created the math gender gap on the test in general also exhibited large differences between white females and males; 10 of the 38 questions which created large score gaps for women of color were only a problem for one group of women, while the remaining 28 were problematic for two or more groups. Two questions made a difference for women in every racial/ethnic group:

Item 25 in Section 2: "If one of the solutions of the equation $x^2 + x + c = 0$ is 2, what is the other solution?
(A) - 3
(B) - 2
(C) 0
(D) 3
(E) It cannot be determined from the information given."

Item 6 in Section 5: "The rectangle above contains two circles, tangent to each other and each tangent to three sides of the rectangle. Which of the following pairs of numbers CANNOT be the length and width, respectively, of the rectangle?
(A) 2, 1
(B) 12, 6
(C) 16, 10
(D) 22, 11
(E) 32, 16

Table 26 shows all the questions on which females in each racial/ethnic group performed worse than males in their racial/ethnic group, compared to female performance on the test in general.

■ **The "Racial/Ethnic Gap":** Virtually no prior research has been published on the differences between female and male performance in any racial/ethnic group other than African Americans, nor are comparisons usually made across the racial/ethnic spectrum (comparing men and women of color to white men and women). Perhaps this lack of research is due to the fact that the gender differences within racial/ethnic groups are so much smaller than the "racial/ethnic gap" between white students and students of color, which has been well documented by ETS researchers (cited below) and others (FairTest). The outstanding exception has been the math performance of Asian Americans; men outperform, and females score almost as well as, white males. Several studies have attempted to identify the major causes of the large score differences between white students and both African American and Hispanic students on the Verbal Sections of the test.

Researchers have found that African American students take longer to finish the test than white students with comparable SAT Verbal scores (Schmitt and Bleistein, 1987; Schmitt and Dorans, 1987). African Americans, like women, perform better when the subject content is about human relations but worse on scientific content questions (Schmitt and Bleistein, 1987). Vocabulary items also cause more problems for African Americans than do reading comprehension sections. Analogies (particularly the easiest ones) and homographs (words with the same spelling but different meanings) also cause more difficulty (Schmitt and Bleistein, 1987); in fact, Schmitt and Bleistein (1987) found that African Americans do less well on analogies because they take longer to finish the test.

Researchers have found that content of interest (which occurs mainly in sentence completion and reading comprehension items) improves the performance of Hispanic and African American students (Schmitt and Dorans, 1987). Hispanic students also perform considerably better on questions that contain words that are true cognates (come from the same root) in Spanish. This especially benefits Puerto Rican and Latin American students, who are more likely to speak Spanish as a second language (Schmitt, Curley, Bleistein and Dorans, 1988; Scheuneman and Briel, 1988). Research has also found that Hispanic students tend to respond to fewer

TABLE 26

Items With Wide Male/Female Variance, by Race/Ethnicity Showing Percentage Differences [D] and Ratios [R]

	White		Black		Asian		Hispan.		N. Am.		Other		All Women	Ratio F/M
	%		%		%		%		%		%			
No.	D	R	D	R	D	R	D	R	D	R	D	R		
VERBAL Section 4														
2	-10		-22				-19		-14		-13		-12	
6									-10					
15										.67				
23							-10							
25	-27	.37	-18	.38	-16	.52	-13	.54	-24	.29	-22	.45	-25	.390
41	-15				-10		-13	.62	-19	.58	-14		-15	
Section 1														
1			-15				-13							
14							-12		-11					
37	-10							.67					-10	
40										.68				
MATH Section 2														
2			-10											
7	-12						-16		-21		-14		-13	
8	-12		-12				-15		-15		-13		-12	
9							-11		-12					
12	-11				-10		-10				-11		-11	
16					-10		-12		-11					
17	-10							.64		.67			-10	.697
18	-10	.66			-11	.67		.58		.68			-10	.643
19	-10													
20					-13	.69				.69				
21	-10	.58			-10	.62								.609
23	-10	.68			-11	.66								.690
24														.688
25		.60		.67		.65		.67		.55		.62		.600
Section 5														
4									-10					
5									-10					
6	-16		-13		-12		-19		-20		-14		-16	
9							-10							
10			-11				-16		-13		-11		-10	
11							-10		-10					
12									-13					
13							-13		-10					
15							-12		-11					
16	-10													
18	-13				-10		-15		-14		-10		-12	
19									-10					
21							-11							
22							-11		-11					
23					-10									
24							-10		-11	.69				
26					-10					.64				
27	-10				-10			.65						
29	-12				-10		-13		-17	.68	-11		-13	
30	-10						-12				-12		-10	
31	-12				-10		-11	.68	-13				-11	
32										.67				
33	-12	.60			-13	.65							-11	.607
35		.50		.67		.54						.57		.429

questions at the end of a section than do whites with comparable SAT-Verbal scores, suggesting that the test's speededness is a problem (Schmitt and Dorans, 1987).

Crouse and Trusheim (1988) found that SAT scores greatly reduce the acceptance of African Americans into all but the least selective colleges; yet the scores were of minimal value in predicting their college performance: "The SAT has very little effect on admissions outcomes over high school rank alone, except insofar as the test lowers Black acceptance" (p. 107).

■ **Questions Showing Large Percentage Differences Between Women of Color and White Women:** The 4,441 African American women in this study performed worse than white females on every question on the test. Over half the Verbal questions (53 out of 85) and 80 percent of the Math questions (49 out of 60) showed differences of more than 10 percent or had large ratio differences. An even greater difference was found in comparing both groups to white men (white women averaged 57 points lower than white men). African American women performed worse on every question, compared to white men, with 71 percent (60 out of 85) of the Verbal and 82 percent of the Math questions showing differences of more than 10 percent.

The 2,373 Hispanic women performed better than white women on 5 of the Verbal questions. On one question Hispanic women performed more than 10 percent better ("the opposite of 'commodious'"), but they were more than 10 percent lower in correct answers or had large ratio differences on almost half the Verbal questions (42 out of 85) and over two-thirds of the Math questions (43 out of 60).

The 601 Native American women found one question considerably easier than did white females ("Rebel:Insurrection"), but they did much worse than white women on 20 Verbal questions and 28 Math questions.

On the other hand, the 2,724 Asian American women performed better than white women on 80 percent of the Math questions; they scored somewhat higher on 42 questions and more than 10 percent higher on 6 questions. They did better on 8 Verbal questions but worse on 24 others.

These data suggest that, with the exception of Asian American women, a large number of questions are causing the score differences between women of color and white women and, further, white men. Appendix F includes all of the questions which had a 10 percent or greater difference in correct answers or a large ratio difference for women of color.

The Gender Gap at the Top: Correlating SAT Scores With High School Performance

Although the main purpose of the SAT is to predict first year college grades, not high school performance, researchers have consistently found that the high school grade point average (GPA) is the best single predictor of college GPA (American College Testing Program, 1973; Breland, 1978; Novick, 1982; R.L.Linn, 1973 and 1982; *ATP Guide*, The College Board, 1988). Correlating SAT scores with current classroom performance for the sample of 1987 test takers, by comparing scores for each sex to their self-reported high school GPAs, therefore should be revealing of the test's predictive ability, as earlier studies have shown high correlations of .7 to .9 between self-reports and corresponding objective measures (Clark and Grandy, 1984).

The sample was divided into four GPA categories: A+ to A; A- to B+; B to B-; C+ to F; there were more girls than boys in each GPA category except the lowest (C+ to F). Not every student who took this test reported a GPA.

However, 90 percent (51,242 of the 54,606 females and 41,742 of the 45,391 males) indicated GPAs on the SAT Student Descriptive Questionnaire, so these data are still representative of the entire group.

In one of this study's most surprising and distressing findings, analysis showed that *the higher the grades, the larger the gender gap.* The biggest sex differences in SAT score averages—much larger than the national averages for the test as a whole—occurred at the highest GPA level (A+ to A), while the smallest gender gap occurred at the lowest GPA level. As Table 27 shows, women with A+ grades averaged 23 points lower on the Verbal Section than men with A+ grades; this is a substantially larger gap than for women in general (14 points). Further, these A+ women averaged 60 points lower than A+ men on the Math Section, compared to 44 points for women in general.

A standard explanation for the larger math gap would be to assert, as a College Board spokesperson often does, that women with A+ grade point averages are more likely to have earned them in English, humanities and language courses while the A+ boys are more likely to have taken courses that prepared them for the SAT-Math, such as physics, chemistry and calculus. However, this explanation fails to account for the larger gender gap on the SAT-Verbal section, where one would expect the high achieving girls with English and humanities backgrounds to excel.

This is one of the most important findings of this study—that the highest achieving girls are penalized the most by the SAT score gap. Their lower SAT scores in comparison to high achieving boys make the test less predictive for them. This may have the effect of excluding these young women from entering the most prestigious colleges that accept their male peers and may also prevent these women from qualifying for merit scholarships and other scholarships that are based on SAT scores rather than high school performance.

Indeed, Federal District Judge John M. Walker (United States District Court for the Southern District of New York) recently enjoined the New York State Department of Education from awarding merit scholarships to high school students based solely on their SAT scores. In his Opinion, Judge Walker wrote that such a use of the SAT discriminates against girls, "in violation of Title IX and the equal protection clause of the U.S. Constitution" (See Appendix I for the full text of the Opinion and Order). The Department of Education may now only use SAT scores as a criterion for scholarship awards in conjunction with high school grades.

The comparison of the percentage of girls and boys answering each question correctly for each of the four GPA groups found that there were more questions with large differences in percentage of correct answers—where girls averaged more than 10 percent lower—between females and males with the highest GPAs than between females and males with the lowest GPAs or between females and males on the test in general. In the Verbal Section, 6 questions showed large sex differences favoring men (compared to 4 for the test in general) and in the Math Section, 22 questions had more than 10 percent differences (compared to 17 Math questions for the test in general), all favoring males; these included the last 5 questions in both Math Sections. The questions are listed in Table 28, indicating which were more difficult for girls with A+ GPAs and for both A+ girls and girls in all GPA categories.

■ **Score Averages by Quartile Compared to GPAs:** Another comparison affirms and highlights this predictive disparity for high achieving females. To establish quartiles, the study put all 100,000 students into rank order and

TABLE 27

Average SAT Scores for Females and Males in Each GPA Category

GPA	Female (Number of Students)	Male (Number of Students)	Difference
Verbal Averages by GPA			
A+ to A	514 (7,492)	537 (5,406)	- 23
A- to B+	457 (17,033)	477 (12,033)	- 20
B to B-	404 (19,387)	425 (15,895)	- 21
C+ to F	363 (7,330)	382 (8,408)	- 19
Math Averages By GPA			
A+ to A	564	624	- 60
A- to B+	495	554	- 59
B to B-	430	481	- 51
C+ to F	378	421	- 43

TABLE 28

Questions That Were More Difficult for Girls Than Boys with A+ GPAs

VERBAL

Questions that were only harder for A+ girls:

Section 1—No. 7, 26, 40

Questions that were harder for girls in all GPA categories:

Section 1—No. 37 (except C+ girls)

Section 4—No. 25, 41

MATH

Questions that were only harder for A+ girls:

Section 2—No. 10, 19, 20, 22

Section 5—No. 23, 25, 26, 27, 32, 34

Questions that were harder for girls in all GPA categories:

Section 2—No. 12, 16 (except C+ girls), 17, 18, 21, 23, 24, 25

Section 5—No. 6, 16, 19, 24 (except C+ girls), 29, 31, 33

divided them into four equal groups based on their Verbal scores and their Math scores: students with SAT-V scores up to 350 and SAT-M scores to 390 were in the lowest quartile; students with SAT-V scores from 351 to 430 and SAT-M scores from 391 to 470 were in the low-mid quartile; students with SAT-V scores from 431 to 500 and SAT-M scores from 471 to 560 were in the mid-high quartile; and students with verbal scores over 500 and math scores over 560 were in the highest quartile.

The comparison of each group's Verbal and Math SAT scores to self-reported high school Grade Point Averages supported the previous finding. Within every SAT score category, girls received higher grades than boys. In the highest verbal quartile, there were 5 percent fewer A+ males than females and in the highest math quartile, there were 10 percent fewer males than females.

TABLE 29

Comparison of GPA to SAT Quartiles by Sex*

Percentage of Girls scoring in each Quartile on Verbal SATs					Percentage of Boys scoring in each Quartile on Verbal SATs			
	Quartiles				Quartiles			
GPA CATEGORY	Low	Low-Mid	Mid-High	High	Low	Low-Mid	Mid-High	High
A+ to A	3	7	16	34	2	5	11	29
A- to B+	19	30	40	43	15	23	34	39
B to B-	48	46	36	20	42	46	40	25
C+ to F	30	17	8	3	41	25	14	6

Percentage of Girls scoring in each Quartile on Math SATs					Percentage of Boys scoring in each Quartile on Math SATs			
	Quartiles				Quartiles			
GPA CATEGORY	Low	Low-Mid	Mid-High	High	Low	Low-Mid	Mid-High	High
A+ to A	2	7	17	39	1	3	8	29
A- to B+	19	32	43	43	11	20	32	41
B to B-	49	46	34	16	43	48	44	25
C+ to F	30	14	6	2	45	28	16	5

*Since 7 percent of the students did not report grades, these quartiles are approximations. But even with all students reporting grades, the conclusions would not change significantly.

Other Explanations

■ **Omission of Questions:** Another critical discovery came from the analysis of the number of women and men who omitted each question (left the answer blank) on the test: women omitted more questions than men by a surprisingly wide margin in both Math Sections. A larger percentage of girls than boys left all but 10 of the 60 Math questions blank; girls' omissions equalled boys' on 9 of the 10 and were lower than boys' on only one question (Section 2/24). An even larger percentage of girls also omitted the last 5 questions in both Verbal Sections and the last 10 questions (except number 24) in both Math Sections. The number of omissions for each question by sex can be found in Appendix G.

Several theories suggest explanations for girls' greater tendency to omit items. Graf and Riddell (1972) found that girls were slower than boys at solving math problems set in a traditional male context; they suggested that "one could significantly decrease between-sex differences in problem-solving by giving power tests rather than tests which rely heavily upon speed."

Research also shows that girls are less likely to be risk-takers and to guess at the right answer, largely because of their different upbringing, socialization, and earlier education (deNys and Wolfe, 1985; Sadker and Sadker, 1985). Linn, DeBenedictis, Delucchi, Harris and Stage (1987) found that 13 to 17 year old girls were more likely to use the "I Don't Know" response on the National Assessment of Educational Progress (NAEP) science assessment, "especially for items with physical science content or masculine themes such as football." They suggest that "an unwillingness to take risks may . . . lead females to avoid giving a definite answer." John D. Miller and Robert Suchner at Northern Illinois University are conducting a "Longitudinal Study of American Youth," using the 1987 7th and 10th Grade National Probability Sample, and have found that gender differences appeared *favoring females* on NAEP math tests when the "I don't know" option was removed. These test results correlated well with the students' 7th and 10th grade classroom performance, where girls were earning higher grades than boys, in contrast to NAEP tests with the "I don't know" option, where girls scored worse than boys. Their research on the NAEP science tests is finding results similar to Linn, *et. al.*

Another conclusion that could be drawn from these studies is that girls may be more likely to follow instructions or "play by the rules." Before each administration of the SAT, the monitor reads the following instructions: "Scores on these tests are based on the number of questions answered correctly minus a fraction (1/4 point per question) of the number of questions answered incorrectly. Therefore, random or haphazard guessing is unlikely to change your scores" (*The Supervisor's Manual*, The College Board, 1988-89). This admonition about guessing—with the information that students are penalized for wrong answers—is probably taken more seriously by girls (it is interesting to note that the "guessing penalty" has been removed from the Graduate Record Examination (GRE) but not from the SAT).

As research on the "I don't know" option shows, girls are more hesitant about guessing when they are not sure of the correct answer, while boys are more willing to guess and probably take the SAT monitor's warning less seriously as well. As Harvard's Carol Gilligan told Rosser in 1987, "this test is a moral issue for girls; they think it is an indication of their intelligence, so they must not cheat. But boys play it like a pinball game."

■ **Time Pressure:** To determine whether either sex was more affected by time pressure, males' and females' performance on the last 10 items on each section of the test were compared to their performance on the rest of the test and to each other. A larger percentage of girls than boys omitted the last 5 questions on Verbal Section 1 (a science reading comprehension passage about a comet) and Verbal Section 4 (analogies). But, large percentages of both sexes omitted questions in the middle of both sections—on analogies, antonyms, and another science reading comprehension passage. In a number of cases, a larger percentage of boys than girls omitted these questions, indicating that content as well as timing was a problem for both sexes.

However, the omissions on the Math Sections told a different story. Both males and females omitted the last 9 questions on Math Section 2 and nearly

all of the last 9 questions on Math Section 5 in larger percentages than for most of the other math questions, indicating that *both* boys and girls ran out of time. But on all but one of these final questions, a larger percentage of girls omitted them than did boys. This indicates that girls have a greater problem with time pressure on the Math Sections of the test than boys do.

Research on differential speededness has been sparse. ETS researchers Wild, Durso and Rubin (1982) studied the effects of increased time on the verbal and math experimental sections of the GRE, also published by ETS, to determine whether increasing the amount of time per question (while controlling for ability) improved the scores of women, African Americans and people returning to college after a number of years out of school. They found that "a larger proportion of examinees complete the experimental tests when given additional time [but] this extra time does not differentially help any of the groups studied." They concluded that the impact of timing on test scores by ability level, particularly within these subgroups, requires further study. Wendler and Carlton (1987) also advise further examination of differences due to speededness, saying that "differential performance may appear . . . at least partially, as a result of test speededness rather than as a reaction to specific item characteristics." And Wing (1981) found that practice effects can be decreased by increasing the time available per question.

As noted earlier, research shows that girls take longer to solve math problems set in male-oriented contexts. According to ETS researchers Lawrence, Curley and McHale (1988), girls also find technical science reading comprehension passages and "true science" sentence completions (as opposed to "surface science"—items whose context could be easily shifted to politics, art or economics) more difficult, suggesting that additional time would be helpful for these questions as well.

It is important to note that this artificial emphasis on speed is the antithesis of the current educational interest in teaching higher level thinking skills. This type of speeded test rewards the facile test taker rather than the sophisticated, thoughtful thinker who gathers new information and organizes, evaluates, and expresses original thoughts clearly and concisely. California State University Professor Arthur Costa, a leader in the field of critical thinking, explains that: "In teaching students to think, the emphasis is not on how many answers they know. Rather, the focus is on how they behave when they *don't* know." Costa suggests that a key measure of a student's growth in intellectual behavior is a decrease in what he calls "impulsive answers": "As students become less impulsive, we can observe them gathering more information before they begin a task, taking time to reflect on an answer before giving it . . . and planning a strategy for solving a problem" (Costa, 1985). Obviously, higher level thinking behaviors such as these cannot be used or tested on a speeded test such as the SAT.

- **Socioeconomic Factors:** This study corroborated other research which has found that social class, measured by parental education and income, was highly correlated with SAT performance for both sexes. However, a gender gap of 49 points or larger remained at every educational and income level.
- **Parents' Education:** The 100,000 students in the sample were separated into six levels of parental education (parents with graduate degrees, parents with some graduate education; parents with bachelors degrees; parents with associate degrees; parents with some college education; and parents with no college education). The comparison between the percentage of correct answers for females and males in each level showed—surprisingly—that higher levels of parental education did *not* narrow the gender gap.

It was expected that females and males whose parents had graduate degrees would perform similarly. However, these girls found as many questions difficult as girls and boys on the test in general. This is consistent with the College Board's findings in 1988 that scores averaged 49 to 63 points lower for females compared to males at every educational level. The gender gap for girls from highly educated families was 56 points. In this sample, four Verbal questions showed clear gender differences favoring boys; 3 were the same questions—Section 1/37, Section 4/25 and 41—that girls in general found more difficult and one was different (Section 1/40). Seventeen math questions showed large gender differences, as they had for girls on the test in general (although they were not always the same questions); these questions were: Section 2/8, 17, 18, 19, 20, 21, 22, 23, 25 and Section 5/6, 20, 26, 27, 29, 31, 32, 33.

However, parental graduate education did correlate with higher scores. Girls whose parents had graduate degrees performed much better than girls whose parents had no college education. Daughters of parents with graduate education did better than daughters of parents with no college education by more than 10 percent on 48 of the 85 Verbal questions and 37 of the 60 Math questions.

■ **Parental Income:** An unexpected finding in this socioeconomic-status cluster came from comparing girls and boys from high income homes. The sample was divided into seven income groups: over $70,000, $60-70,000, $50-60,000, $40-50,000, $25-30,000; and lower than $25,000. Although a smaller number of questions showed large differences between the sexes at the highest and lowest income levels, the score gap between males and females remained large at every income level. For the students from homes with incomes over $70,000, only 8 questions showed large sex differences—3 on the Verbal Section and 5 on the Math Sections (Verbal Section 4/4, 25, 41; Math Section 2/21, 23; Math Section 5/6, 23, 33). For the lowest income group, 3 Verbal questions and 9 Math questions showed large sex differences (Verbal Section 4/2, 25, 41; Math Section 2/7, 8, 11; Math Section 5/7, 10, 18, 29, 30, 31).

Parental income correlated with high performance in a predictable way when girls from the highest income families (over $70,000) were compared to girls from the lowest income families (less than $25,000). The wealthiest girls did considerably better on 24 Verbal questions and 32 Math questions than girls from the lowest income families.

This is not surprising. In 1980, Allan Nairn and Ralph Nader charged in *The Reign of ETS: The Corporation that Makes Up Minds* that family income correlates so highly with SAT scores that the scores are "class in the guise of merit" (p. 204). ETS denied this, saying that although "average scores are higher for students from families with higher incomes, students from each income level obtain the full range of SAT scores" (Crouse and Trusheim, 1988). Crouse and Trusheim found a National Longitudinal Study (NCES) of students applying to four-year colleges which showed that "average family income rises with each 100-point increase in SAT scores, except for the highest category where the number of cases is small" (1988, p. 126).

However, the income picture for girls is different. Although SAT scores rise with family income level, there is still a high income gender gap; girls from the highest income families (over $70,000) receive lower average scores than boys at this income level. In fact, highest income girls' math score averages are the same as those of boys from the middle income range ($40-50,000); and their verbal score averages are the same as those of boys who are less affluent ($60-70,000 range). This means that class does not predict

SAT scores for girls the way it does for boys. Indeed, when ETS suggests that the larger number of low income girls compared to boys are pulling the female averages down, it is ignoring the fact that girls at every income level score worse than boys with comparable family incomes (see Appendix H).

■ **Questioning the Value of the SAT:** The test publisher is well aware of the SAT's underprediction for women. As ETS researchers Clark and Grandy (1984) state, "the underprediction of women's first year college grades has been reported consistently in the research literature." The last page of the College Board's *Admissions Testing Program Guide* for 1987-1988 states that "the validity of high school record is typically somewhat higher than the validity of the optimally weighted combination of SAT scores." The *Guide* reports that for first year women students, the median correlations for high school record and for the optimally weighted combination of SAT scores were .50 and .46, respectively. This raises the question with which Rosser and others began their research: If high school grades have a higher correlation with first year college performance than the SAT, why is the test necessary? As ETS researcher William Angoff says, "past achievement is always a good predictor of future achievement, often a better predictor than aptitude" (Angoff, 1988).

A growing body of research suggests that SAT scores contribute practically nothing to prediction of first year college grades. Crouse and Trusheim, in their book *The Case Against the SAT*, make a statistically compelling argument against the use of SAT scores by colleges and universities, showing that the SAT does not help them improve their admissions decisions. Nor do SAT scores help students to select colleges where they will be successful. Crouse and Trusheim show that SAT scores increase the prediction of future performance by approximately 0.035 of a grade point. The New York Public Interest Research Group has also published a study on test validity and prediction, entitled *Rolling Loaded Dice*, in which half the 20 colleges which have filed predictive validity studies with the New York State Department of Education showed "the SAT Verbal scores could predict grades no better than *four percent above pure chance*—in short, a virtually meaningless statistic for New York college admissions officers." This research must be borne in mind as we consider the implications of sex and race bias on these standardized tests.

Is It Possible to Create a Sex-Fair Test?

by

Phyllis Rosser and James W. Loewen

Construction of Sex-Biased and Sex-Equal Verbal Tests: The existence of verbal SAT items that markedly favor one sex or the other on the June 1986 SAT (analyzed by James Loewen) indicates that the 10 point "gender gap" suffered by girls nationally is manipulable by the content of the included items. Test-makers could easily construct a test on which one sex nationally scored as much as 50 points better than the other. On the June 1986 SAT, for example, if the 10 items that favored boys the most were deleted and replaced with items similar to the 10 items that most favored girls, girls nationally would outperform boys by about 4 points. This change would be accomplished solely with items that could pass through ETS's current screening process.

Thus, "balance" has primarily a political, not intellectual, definition. ETS has long known and stated that "categories designated 'world of practical affairs' and 'science' are typically easier for males, whereas the categories designated 'aesthetics/philosophy' and 'human relationships' are easier for females." ETS apparently believes that its changes in the Verbal SAT, which substituted a male advantage for the previous female advantage, are "balanced" and "seem to accomplish their purpose" (Donlon, 1984, p. 52).

But, since any difference between boys' and girls' means is dependent upon inclusion or exclusion of questions favoring one sex or the other, it is doubtful that the observed national 10 point difference can be considered "real" or that the test that created this difference can be considered "balanced." Instead, items could be included so that no difference in group means for boys and girls would result. As ETS studies the performance of subgroups, items that particularly favor males, whites, and the affluent should be removed or balanced with items favoring females, people of color, and the working-class.

■ **Construction of Sex-Biased and Sex-Equal Math Tests:** As with the Verbal test, averages for males and females can be altered if existing math items favoring boys are replaced by items similar to current items that favored girls. Because boys outscored girls on most math items, a sex-equal math test cannot be constructed solely from existing questions. On the Math SAT nationally, boys now outscore girls by about 47 points on ETS's 200-800 point scale. Since the difference between boys' and girls' means is partly derived from questions favoring males by margins of more than 10 percent, at least 3 of which contained overtly "pro-boy" verbal content, all of this difference cannot be "real." If the 10 most "pro-boy" items were replaced with items similar to the 10 most "pro-girl" items, boys nationally would outscore girls by about 29 points. Thus, more than a third of the existing math "gap" suffered by girls nationally would be eliminated by excising these 10 items.

Only one math item had any verbal content related to girls, and that consisted solely of the proper noun "Judy" in item 11 in Section 2: "Judy doubles k, adds 12 . . ." Otherwise, that item too was gender-free and girls did rather well in solving it, only .5 percent below boys. In contrast, on two items set in a boys' camp, boys outperformed girls by 12.3 percent and 15.6 percent. And the largest sex-related difference of all—27 percent—appeared on the item dealing with basketball team statistics.

Because the SAT math gap is not replicated in school performance, and because the verbal content of math questions influenced scores by sex, it is clear that ETS could revise its math questions to insert verbal content that overtly includes girls' subculture and female names and omits boys' subculture and male names—just the reverse of current practice on this SAT (with the single exception of "Judy"). This might lead to a further increment of perhaps 5 points in girls' scores, relative to boys' (cf. Donlon, *et al.*, 1977). Moreover, adding items with female verbal content might create items with "pro-girl" differences, which the test does not now contain, thus responding to the findings of the several studies that have shown that females perform better on questions that refer to females or whose content reflects their cultural experience (Donlon, Ekstrom, and Lockheed, 1979; Dwyer, 1979; Stricker, 1982).[10]

A similar attempt has been made to equalize the November, 1987 SAT. The 4 questions favoring boys with the largest percentage differences between the sexes were removed from both the Verbal and Math sections of the test. These were Verbal questions 37 in Section 1 and questions 2, 25, and 41 in Section 4 and Math questions 7 and 8 in Section 2 and questions 6, 18, and 29 in Section 5.

Raw scores were recalculated to determine whether removing these questions with large gender differences would appreciably reduce the SAT gender gap. Although this made a difference on the Verbal Section, it did not affect the scores on the Math Section. Boys' and girls' correct verbal averages changed from a difference of 2 questions favoring males to boys and girls answering an equal number of questions (41) correctly. However, girls averaged 2 more wrong answers (28) than boys (26), reducing their total score to 34 compared to 34.5 (equalling 35) for boys (see Tables 30 and 31 below).

When the 4 questions with the largest gender gap were removed from the Math Section, girls still averaged 3 fewer correct answers (28) than boys (31). Girls also averaged 2 more wrong answers than boys, reducing their raw score to 22.25 (equalling 22) compared to boys' raw score of 26.83 (equalling 27). The gender gap remained the same as before these 4 questions were removed—5 raw score points difference between the sexes. However, by only eliminating 4 items and not replacing them with items favoring women, we had not expected to make a significant difference on the Math test.

Overall, these findings confirm the earlier analysis of the June 1986 SAT conducted by Loewen and further support the contention that ETS could easily construct a sex-equal Verbal test simply by including a few more questions set in the context of experiences more familiar to females and eliminating a few of the questions that are most clearly set in a context familiar and comfortable to males. Since ETS tests all questions on the

TABLE 30

November 1987 SAT Results With No Questions Removed

	raw score calculation	RAW	SAT
VERBAL			
Girls	42 - 1/4 (29) = 34.75	35	430
Boys	44 - 1/4 (27) = 37.25	37	440
MATH			
Girls	31 - 1/4 (12) - 1/3 (6) =	26	460
Boys	35 - 1/4 (11) - 1/3 (5) =	31	510
TOTAL			
Girls		61	890
Boys		68	950

TABLE 31

November 1987 SAT Results With Four "Worst" Questions Removed From Each Section [Raw Scores Only]

	raw score calculation	RAW
VERBAL		
Girls	41 - 1/4 (28) = 34	34
Boys	41 - 1/4 (26) = 34.5	35
MATH		
Girls	28 - 1/4 (11) - 1/3 (6) = 22.25	22
Boys	31 - 1/4 (10) - 1/3 (5) = 26.83	27
TOTAL		
Girls		56
Boys		62

experimental sections of the test before using them, it should not be difficult to balance the Verbal Section. Equalizing the Math Section, however, appears to be more complex; extensive additional research may be needed to determine how this test can be made fairer to women and more predictive of their first year college performance.[11]

High School Achievement Tests—Are They Fair for Girls?

Most high schools across the country administer standardized achievement tests to students at each grade level to measure their progress and to evaluate schools' performance. These tests include the California Achievement Tests and Comprehensive Tests of Basic Skills published by CTB/McGraw-Hill; the Metropolitan Achievement Tests published by The Psychological Corporation; the Iowa Tests of Basic Skills published by The Riverside Publishing Company; and the Sequential Tests of Education Progress (STEP) and School and College Ability Tests published by Educational Testing Service. They provide two types of data: how well a student performs compared to students nationally at each grade level (a norm-referenced interpretation); and, how well a student has learned a particular skill (a criterion-referenced interpretation). Each set of tests is tried out on thousands of students from different socioeconomic and racial groups to determine national score averages, or norms.

In 1987, the norming of the six major achievement tests was questioned by Friends for Education, a citizens group working for accountability in education. A survey conducted by its president, physician John Cannell, found that over 90 percent of the nation's school districts, and over 70 percent of the nation's students had "median" scores above the national 50th percentile. Even in districts or states where scores would be expected to be low based on other measures (including education expenditures, for instance), such as Alabama, Georgia and Mississippi, scores were not low in comparison to national norms. Cannell attributes this apparent improved performance to "inacccurate initial norms" and teaching to the test (Cannell, 1987). No sex or race differences were mentioned in the norming process and districts with large urban populations (such as Trenton, New Jersey and New York City, for example) claimed to be above the national average (Cannell, 1987).

The technical reports for these achievement tests, provided by the various publishers, include considerable demographic information on the populations used for standardization (norming). Some list percentages of African American and Hispanic students enrolled in the norming school as well as percentages of special education students. However, only one technical report listed the number of males and females participating in the norming process: the *California Achievement Tests: Form E and F Technical Report.* Means and standard deviations, as well as the number of questions biased for and against males and females, were given for each subtest. The same data also were provided for Asian Americans, African Americans and Hispanics, by sex. According to CTB/McGraw-Hill product manager John Stewart, item analyses are conducted on questions during the tryout phase, before the test is standardized. In a telephone conversation with Rosser (January 11, 1989), Stewart said that "very little bias was found on the California Achievement Test and those questions were balanced so that an

equal number of items favored each sex." Questions also were analyzed by sex and race with a norming sample of African Americans and Hispanics in the same number or a greater percentage than their representation in the population in general.

■ **Girls' Score Averages Are Higher than Boys' on the Major Standardized Achievement Tests Used in High School:** Female/male performance differences on the California Achievement Test have also been studied extensively by Donald Ross Green, CTB/McGraw-Hill's Manager of Basic Research. Using the 1985 standardization data, he looked at a representative sample of 110,000 students in grades K-12 who took 72 of the basic batteries (33 additional batteries were not studied). Green found that girls scored consistently higher than boys on most of the tests—in all ethnic groups examined (white, African American and Hispanic). Girls' higher performance resulted from better performance on almost all test items, rather than from a small group of items, while boys' performance tended to be more variable than girls', for all ethnic groups studied.

The number of items biased in favor of females (females were expected to perform better than males) was less than the number of items biased in favor of males—and these items made up less than 10 percent of the test. The questions were judged to be biased because of differential familiarity with the content or identification with the sex of the principal people in the questions. Green says that "it was surprising to me that an item can be biased merely because of the sex of the person described in this item," supporting the 1979 findings of ETS researchers Ekstrom, Lockheed, and Donlon. Green speculates that language differences between males and females, found by other researchers to create a difference in item performance, could be a factor (Green, 1987).

In 1978, Plake, Hoover, and Loyd examined the Mathematics Problem Solving (MPS) and Mathematics Concepts (MC) of the Iowa Test of Basic Skills at grades 3, 6, and 8. They found that problem solving was more difficult than concepts for students at all three grade levels but that girls performed better than boys on the tests overall—a finding that contradicts girls' lower math performance on the SAT over the past 25 years.

■ **A Look At One High School's Experience:** To see first hand whether standardized achievement test scores correlate with classroom performance and to assess the predictive validity of standardized tests, Rosser conducted a study of 203 high school seniors in the Class of 1988 at Holmdel High School in suburban New Jersey, comparing their English and Math grades to their English and Math CTBS Achievement Test scores. For the 102 females and 101 males, she found a very high correlation between CTBS Reading scores and English grades of .70 for the girls and an excellent correlation of .47 for the boys. The correlation between math scores and math grades was also good, although it was much lower for females than the English: .45 for the boys and .43 for the girls.

For these students, the CTBS Achievement Test appeared to be an excellent predictor of classroom performance and did not seem to penalize girls in any way. However, these students had even higher correlations between their SAT Verbal and Math scores and their English and Math grades (see Table 32 below) suggesting a level of mathematics preparation that was considerably above the norm. In this regard, it may be important to note that a number of these students come from homes in which one or both parents are scientists.

TABLE 32

Correlations Between Grades and Test Scores for 203 High School Seniors

	Females	Males
Class Rank to Total SAT Score	.72	.63
English Grade to CTBS Reading Score	.47	.70
Math Grade to CTBS Math Score	.45	.43
English Grade to SAT Verbal Score	.63	.56
Math Grade to SAT Math Score	.57	.55
CTBS Reading Score to SAT Verbal Score	.62	.59
CTBS Math Score to SAT Math Score	.61	.54

■ **Longitudinal Studies—Cause for Concern:** In light of this achievement test data, the findings of two recent national longitudinal studies of high school performance are cause for concern. Conducted by the National Center for Education Statistics (NCES) and the federally funded National Assessment of Educational Progress (NAEP), they show deficits in female performance similar to those in the SAT. These results raise questions about political intent; both studies used tests written by Educational Testing Service and these findings are often cited by ETS researchers to justify the gender gap on the SAT.

The NCES study was conducted by ETS researchers in 1985 and used statistics from two national tests written by ETS—the National Longitudinal Study (NLS) administered to high school sophomores and seniors in 1972 and the "High School and Beyond" (HSB) administered in 1980—to document changes in academic achievement. The 1980 results for 28,240 (51.4 percent female and 48.6 percent male) high school seniors showed that both women and men had declined in Reading Comprehension and Vocabulary score averages. But the women had lost their 1972 lead over men in Reading Comprehension and were now performing about the same. In Vocabulary, female score averages declined so much that the males were outperforming them (Mullis, 1987).

"High School and Beyond" also included a follow up study of the 1980 sophomores, who were retested in 1982; at that time, researchers found that the men were performing slightly better than the women in Reading Comprehension as well as in Vocabulary (echoing the decreasing female SAT Verbal scores). However, female performance was superior to males in the HSB writing tests in both 1980 and 1982 (Mullis, 1987).

All achievement tests *except* HSB show girls outperforming boys in reading from age 9 onward, but as they get older the achievement gap narrows. The NAEP studies have assessed the educational achievement of 9, 13, and 17 year old students in 1970-71; 1974-75; 1979-80; and 1983-84. Approximately 22,200 students were tested for Reading Proficiency at each

age level, with nearly equal numbers of males and females (except in the 17 year-old sample, which was 51.3 percent male and 48.7 percent female). The NAEP studies found that girls' reading proficiency at all three ages was declining in the 1980s, while boys made steady gains, narrowing the reading proficiency gap (Mullis, 1987). This is particularly troubling, as Reading is an area in which girls traditionally have received higher scores, but in 1984, 34.8 percent of 17 year-old boys and 43.9 percent of 17 year-old girls were "adept" readers compared to 31.5 percent of boys and 42.7 percent of girls in 1971.

NAEP assessments of mathematics found few sex differences at ages 9 and 13, but males outperformed females at age 17, even when general course background was held constant; in HSB math tests, males outperformed females as sophomores and seniors, but women had higher average math grades—even in advanced math courses (Klein, 1986). However, like "High School and Beyond," girls clearly performed better than boys on NAEP's writing assessments, with no changes in the size of the differences between the sexes from 1979 to 1984 (Mullis, 1987).

Other achievement test trends appear more ominous. In 1986, state-wide testing of high school juniors in Maine found large gender differences, with boys outperforming girls in math (283-218) and significantly outperforming them in science (314-187) and social studies (304-197). Girls outscored boys in reading (288-213) and the humanities (266-234) and significantly outscored them in writing (298-203) and writing mechanics (325-174). Assessment Director Paul R. Walker proclaimed the assessment an "unqualified success" but found the difference between the sexes "most startling" (Portland *Press Herald*, September 19, 1986). Again, researchers should question the purpose of achievement assessments that do not correlate with girls' superior classroom performance in math, science and social studies.

■ **The Narrowing of Cognitive Differences:** Sex stereotypical differences such as the ones found in the Maine assessments are currently being countered by other studies that show a narrowing of cognitive differences between the sexes. Yale Professor Alan Feingold examined normative data for the PSAT collected between 1960 and 1984 and for the Differential Aptitude Test (DAT) between 1947 and 1980 and found that gender differences had declined "precipitously" over the years on both tests. The important exception was the "well-documented gender gap at the upper levels of performance on high school mathematics which has remained constant over the past 27 years" (Feingold, 1988, p.95).

Two important meta-analyses by Janet Shibley Hyde and Marcia C. Linn have also found cognitive gender differences disappearing. In 1988, Hyde and Linn analyzed 165 studies of gender differences in verbal ability and found differences in favor of females so small that they could "effectively be considered to be zero" (Hyde and Linn, 1988). The one outstanding exception was female performance on the SAT-Verbal, where the gender difference has been increasing. In their 1988 meta-analysis of gender differences in mathematics, which has yet to be published, they also found that math differences between the sexes were small. The largest differences were found on questions that drew on advanced coursework in math and were similar to the gender differences in course enrollment for these subjects. Since differences on most national assessments were declining, Linn and Hyde suggest that the "large, consistent gender differences found for the voluntary SAT-M sample are anomalous" (Linn and Hyde, 1988).

The evidence that achievement tests predict classroom grades equitably for both sexes is conflicting but these test results appear to be less damaging

to girls' educational opportunities than the SAT, PSAT or ACT. It is not clear why girls find standardized achievement tests administered at high school grade levels less difficult but they seem to show that multiple choice tests are not *a priori* more difficult for females.

Perhaps one explanation can be found in the different purposes and premises of "achievement" tests as compared to "aptitude" tests. High school achievement tests measure information learned during exposure to a particular subject, while "aptitude" tests draw on ability developed from a wide range of human experience (Angoff, 1988). Indeed, in a recent article in *American Psychologist* (September 1988), ETS researcher William Angoff stated that "aptitudes are frequently in continuous change and therefore cannot be innate . . . [they] are indeed susceptible to differential cognitive training." It is difficult to understand how English and math "aptitude" so defined can be separated from classroom exposure and achievement in these subjects; nor is it clear that the SAT's purported assessment of "aptitude" adds significant information to the knowledge of students' abilities and achievements reflected in their grades.

Review of the Literature on Gender Bias in College Entrance Examinations[12]

The study of gender bias in college entrance examinations is closely tied to the study of sex discrimination in general as well as to the study of all psychometric bias. A bibliography that attempted to include all of these related works, however, would be massive and of little use to those researchers interested in the specific issue of gender bias in college entrance examinations. Therefore, this review and the complete bibliography found at the end of this report include works that either contain direct references to the SAT, ACT, or achievement tests; refer to the issue of gender bias with regard to widely used basic skills tests administered to high school students; or focus on broader or related issues in ways that are immediately relevant to the study of gender bias in college entrance examinations.

In addition to this brief overview of the literature, 19 annotated entries are highlighted because, for better or worse, they have defined and refined the issues pertinent to the study of bias in testing. The research questions asked by many of these studies address gender differences on college entrance examinations in the context of well-established debates in psychometrics: Are test score differences indicative of actual differences between males and females, or are they artifacts of the test itself? Assuming that test scores tell us something about actual differences in ability, do these differences have biological or sociological origins? What should society's response be to tests that are found to have an adverse impact on women and persons of color?

Responses to this final question in the literature may range from "nothing" to "regulate the use of the tests to ensure that they do not lead to discrimination."

■ **Do Test Score Differences Tell Us About People Or Tests?:** Julian Stanley and Camilla Benbow are the chief advocates of the argument that SAT-Math score differences reflect actual differences between males' and females' math talents, and are not artifacts of the test (Benbow and Stanley, 1980; 1983a; 1983b). Their longitudinal studies of seventh- and eighth-grade students who achieved high scores on the SAT reveal that these students went on to excel in their high school and college math classes (Benbow and Stanley, 1982). Virtually all researchers at ETS, which produces the SAT, claim that the test scores have significant predictive validity for first year college grades, regardless of the students' gender, and that the tests therefore are fair (Clark and Grandy, 1984). In addition, by first matching groups by their scores, ETS is able to spot individual items that are biased without reducing the ability of the test to indicate areas of genuine differences (Kulick and Dorans, 1984). Thus, ETS argues that the SAT has validity for both prediction and internal consistency.

Others point out that students' test performances are sensitive to factors specific to test-taking situations. Speededness (Donlon, 1977) and anxiety (Payne, 1984; Fyans, 1979; Wildemuth, 1977), for instance, are both related to test environments and test score differences, although they are not intrinsic to learning *per se*. These factors work against women as compared to men (Plake, Ansorge, Parker, and Lowry, 1982; Billingham, 1981).

Some SAT items favor males and some favor females (Milton, 1958; Donlon, 1973; Dwyer, 1976a, 1976b; Lueptow, 1980; Loewen, Rosser and Katzman, 1988). In fact, ETS acknowledges that it has made a number of changes in the SAT over the years that have increased the proportion of items that favor males; this has led some to charge that it is these policy decisions about the test, and not males' academic superiority over females, that allows men to score higher than women (Dwyer, 1976a, 1976b). Still others dispute the ETS claim that the SAT does indeed predict college performance accurately, holding that the under and over prediction problems related to gender are quite pronounced (Rosser, 1987; Hogrebe, 1983; Holland and Nichols, 1964).

■ **Nature or Nurture?:** Among the psychologists who claim that the SAT captures genuine differences in quantitative and verbal skills, the origin of these differences is a major research concern. Many ascribe the differences to socialization, including differential course-taking, parental expectations, and motivation (Adams, 1986; Doolittle, 1985; Clark and Grandy, 1984; Schofield, 1982; Hoffman and Maier, 1966). Others attribute the differences to biology (Stanley and Benbow, 1983).

■ **Title IX and Gender Bias in Standardized Tests:** If sex discrimination results from the use of standardized test scores, women and girls may have recourse under Title IX of the Education Amendments of 1972, which prohibits educational institutions that receive federal funds from discriminating on the basis of sex. The Title IX implementing regulations include discriminatory admissions tests under this prohibition. Lawyers and policy analysts have interpreted these requirements and the recourse available for those who believe they have been discriminated against on the basis of sex biased tests (Fitzgerald and Fisher, 1974; Lockheed, 1974a).

ETS and the College Board, however, maintain that the SAT's accuracy and fairness depend on the ability of psychometricians to design the test in accordance with scientific standards without regard for civil rights laws and

policies; they suggest that interference from courts or legislatures will only reduce the usefulness and validity of the SAT (Anrig, 1987).

The recent federal District Court decision in New York invalidated the New York Department of Education's use of SAT scores as the sole basis for awarding merit scholarships for college. Judge John M. Walker's Opinion states that this method discriminates against girls: "After a careful review of the evidence, this court concludes that SAT scores capture a student's academic achievement no more than a student's yearbook photograph captures the full range of her experiences in high school" (see Appendix I for the full text of Judge Walker's Opinion and Order). This is the first case[13] that has challenged the use of the SAT based on both the Fourteenth Amendment to the Constitution and Title IX and may set an important precedent.

Selected Studies

Benbow, Camilla and Julian C. Stanley. 1982. "Consequences in High School and College of Sex Differences in Mathematical Reasoning Ability: A Longitudinal Perspective." *American Educational Research Journal* 19 (Winter): 598-622.

Benbow and Stanley conducted a five-year longitudinal study, beginning in 1972-74, of 1,996 7th and 8th graders who scored as well as a national sample of 11th and 12th grade males and females, to assess the persistence of sex differences in mathematics achievement. The study found that males as a group scored higher than females on standardized tests, while females received higher grades in their mathematics classes than males with the same SAT scores. The authors conclude that a biologically-determined aptitude for math is greater in males than in females.

Chipman, Susan F. 1988. Word Problems: Where Test Bias Creeps In. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Chipman's research explores the possibility that setting math word problems in contexts familiar to women improves their performance; she reviews other research that supports that conclusion. She also questions the current tendency to consider that "mathematical reasoning ability" is innate and suggests that training in underlying mathematical structures needs to be improved for everyone in the United States.

Clark, Mary Jo, and Jerilee Grandy. 1984. *Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers.* Report No. 84-8. New York: College Entrance Examination Board.

These ETS researchers compare the SAT score prediction differences between males and females for first year college grades, which is the construction criterion for the SAT. Their findings show that the SAT overpredicts first year grades for men and underpredicts first year grades for women. Throughout the report the authors assume the overall validity of the SAT's construction, which is problematic given the initial findings of the test's prediction differentials. Instead, researchers hypothesize that grades mean different things for girls and boys and that boys, for instance, take more "difficult" math courses in high school and college.

Diamond, Esther E. 1985. Content, Context and Construct Considerations in Sex Bias in Testing. Paper presented at the annual meeting of the American Educational Research Association, 31 March-4 April, Chicago.

Diamond provides a thorough review of test construction and sex bias literature; she defines facial, content, context, and construct bias and describes how they occur in standardized tests. She also summarizes the research literature on how each type of bias affects female test performance and presents a variety of issues to be considered in efforts to minimize item bias.

Diamond, Esther E. and Carol Kehr Tittle. 1985. "Sex Equity in Testing." in Klein, Susan, ed. *Handbook for Achieving Sex Equity Through Education.* Baltimore: Johns Hopkins University Press.

Diamond and Tittle synthesize the work of researchers who have assessed sex bias on the major types of educational and psychological tests administered to women and girls. They include a number of recommendations for test use to enhance equity and pose research questions which require further investigation.

Donlon, Thomas F. 1973. *Content Factors in Sex Differences on Test Questions.* Research Memorandum 73-28, Princeton: Educational Testing Service.

Data for this study are not current (SAT item-analyses from 1964), but the findings pointed the way toward the kinds of questions ETS is now asking about the SAT's content formula. Donlon found that males and females did better on different types of math questions, and concludes that if the content of the Math Section of the SAT were limited to algebra, the mean differences could be reduced by 20 points.

Doolittle, Allen E. 1985. Understanding Differential Item Performance as a Consequence of Gender Differences in Academic Background. Paper presented at the annual meeting of the American Educational Research Association, 31 March-4 April, Chicago.

After controlling for math background (level of math), Doolittle finds that girls who take the ACT-Math do better on algorithmic, calculation-oriented items, while boys do better on geometry and word problems.

Dorans, Neil J., and Edward Kulick. 1983. *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach.* Princeton: Educational Testing Service.

Dorans and Kulick advocate the Mantel-Haenzel technique for elimination of item bias because other methods "exhibited undesirable sensitivities to differences in overall subpopulation ability distributions for males and females." The authors hold that the mean gender differential on the SAT-Math Section is "reflective of the difference between the mathematical ability distributions for males and females."

Dwyer, Carol A. 1976. "Test Content and Sex Differences in Reading." *The Reading Teacher,* 29 (8): 753-77.

In an early study, Dwyer discusses the balancing of the verbal content on the SAT so that by the earlier 1970s both sexes received similar average Verbal scores. She questions the justification for this since a large number of studies of verbal tests continued to show sex differences. She also notes that similar efforts were not made to balance the content of the Math Section "even though sex differences have favored males by a great number of points since the first administrations of the test."

Ekstrom, Ruth B., Marlaine E. Lockheed, and Thomas F. Donlon. 1979. "Sex Differences and Sex Bias in Test Content." *Educational Horizons* 58 (1):47-52.

These ETS researchers present findings from item-analyses of the Metropolitan Achievement Test, the Iowa Test of Basic Skills, and the Sequential Tests of Education Progress. The study's results show that girls

do significantly better on questions that are neutral or have female actors, but the authors cannot explain why some items show these differences and not others.

Fallows, James. 1980. "The Tests and the Brightest: How Fair Are the College Boards?" *Journal of the National Association of College Admissions Counselors*, 24 (3):14-31.

First published in the *Atlantic Monthly*, this well-researched article provides a sophisticated overview of the public policy questions raised by the widespread use of the SAT. Fallows is particularly good at making explicit the assumptions on which the SAT is constructed and packaged.

Hyde, Janet Shibley and Marcia C. Linn. 1988. "Gender Differences in Verbal Ability: A Meta-Analysis." *Psychological Bulletin* 104 (1): 53-69.

Hyde and Linn examined 165 studies that reported data on gender differences in verbal ability and found that the differences were "insubstantial." They conclude that gender differences in verbal ability no longer exist.

Loewen, James W., Phyllis Rosser, and John Katzman. 1988. Gender Bias in SAT Items. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

An item analysis of gender differences on the SAT shows that 7 of the 85 verbal items and 10 of the 60 math items favored one sex by more than ten percent. Authors studied the results of a practice SAT taken by 1,112 students in Princeton Review during the second session of their coaching classes, "under conditions as similar as possible to ETS test centers." The item showing the most favoritism to men was a math question that asked about a "basketball team won/loss record"; 27 percent more boys than girls answered that question correctly. Answers were also correlated with self-reported grades, number of courses taken, estimates of test anxiety, and socioeconomic factors. The tests underpredicted females' high school grades, which averaged higher than males' grades in both English and math. Girls' poorer test performance was not linked to test anxiety or time pressure.

Ramist, Leonard and Solomon Arbeiter. 1986. *Profiles, College-Bound Seniors, 1985*. New York: College Entrance Examination Board.

The College Board periodically publishes the self-reported data of seniors who took the SAT. Group scores are identified by quartiles and means. Recent findings indicate that after holding constant such socioeconomic information as income, school attended and parental education, males have a higher mean SAT score than females.

Rosser, Phyllis. 1987. *Sex Bias in College Admissions Tests: Why Women Lose Out*. Cambridge, Massachusetts: FairTest.

This report provides a thorough review of the different ways the SAT is used and how, in each case, females are adversely affected. The Educational Testing Service (ETS) claims that the SAT is designed to predict first-year college GPAs. Rosser's central argument is that since SAT scores overpredict males' GPA during their first year in college, while they underpredict females' GPA during this same period, the tests are unfair. "If the SAT predicted equally well for both sexes, girls would score about 20 points higher than boys, not 61 points lower." Rosser contends that the prediction differential makes the SAT an inappropriate test for college admissions, as well as for the allocation of other educational opportunities, such as scholarships and participation in "gifted and talented" summer programs.

Selkow, Paula. 1985. *Assessing Sex Bias in Testing: A Review of the Issues and Evaluation of 74 Psychological and Educational Tests.* Westport: Greenwood Press.

Selkow evaluates sex bias on psychological and education tests as defined by a number of objectively identifiable criteria, including: stereotypical representations of each gender in illustrations; number of mentions of each gender in questions and answers; number of famous men and women depicted. Results show that males outnumber females in all categories.

Stanley, Julian, and Camilla Benbow. 1983. "SMPY's First Decade: Ten Years of Posing Problems and Solving Them." *Journal of Special Education* 17 (1):11-25.

Stanley and Benbow offer a historical account of their Study of Mathematically Precocious Youth (SMPY), which began in 1971. Researchers administer the SAT to 7th and 8th graders in order to screen for "gifted" students who then could be singled out for special curricula and summer programs. Substantial sex differences in scores on the Math section of the SAT are found: At a score of 500, the ratio of males to females is 2:1; at 600 the ratio of males to females is 4:1; at 700 the ratio of males to females is 15:1.

Tittle, Carol Kehr. 1978. *Sex Bias in Testing: A Review with Policy Recommendations.* San Francisco: Women's Educational Equity Communications Network, Far West Laboratory.

Tittle provides an early, comprehensive review of the types of facial bias (number of females, stereotyped representations, etc.) that occurred in major educational and psychological tests during the early 1970s. She provides an excellent starting point for understanding the issues involved in gender bias in standardized testing.

Wild, Cheryl L. and Carol A. Dwyer. 1980. "Sex Bias in Selection." *Psychometrics for Educational Debate,* L.J. van der Kamp, N.M. de Gruijter and W.F. Langerak, Eds. New York: Wiley.

Wild and Dwyer argue that sex bias models must evaluate both predictor and criterion variables, suggesting, for example, that ACT scores that underpredict female college performance may indicate that the "GPA reliability is less than the reliability of the predictors." The study shows that "only" 12 of 90 ACT verbal questions and 7 of the 60 ACT mathematics questions indicated bias.

Notes

1. The research upon which this section is based was conducted by James W. Loewen (Department of Sociology, University of Vermont, Burlington), Phyllis Rosser, and John Katzman (Princeton Review). James Loewen was the principal author of this study, an earlier version of which was presented at the annual meeting of the American Educational Research Association in New Orleans, LA on April 5, 1988.

2. ETS also relies on self-reported data for its analyses; studies have found rather high correlations (.7 to .9) between self-reports and corresponding objective measures (Clark and Grandy, 1984).

3. Owing to slightly different procedures for determining sex of student on different computer runs, *n*'s and percentages can vary slightly from table to table.

4. This study did not examine the TSWE (Test of Standard Written English). When computing the sample's math scores, we assumed all math items had 5 alternatives. Some have 4, so this procedure slightly under-subtracts for wrong answers, giving the

students slightly (less than 10 points) higher math scores than they should have. ETS uses an irregular scale to convert raw scores to SAT scale scores; hence, getting one more item correct can increase SAT scores by 0, 10, or 20 points. We converted the ETS scale to regular intervals.

5. As the overall mean differences imply, girls and boys group about the same in verbal scores, while 11 percent more girls fall in the low math group, compared to boys. Thus, the overall male/female difference in math exam performance is greater than Table 4 displays, because more males than females fall in the higher columns of the table.

6. r between the V-SAT and high school GPA = .28, while r between the M-SAT and high school GPA = .33. These r's are similar to the r's of .3 between SAT scores and first-year college grades reported by Schrader (1984), but lower than the r's of .5 between SAT scores and high school rank in class in Schrader's national study.

7. Of all issues studied, anxiety versus performance was probably most affected by this sample's test conditions. Students' performance on this SAT did not "count." On an administration of the SAT upon which college entrance and scholarships depend, anxiety might affect performance more. Also, boys may not admit as much test anxiety as girls, but may actually feel as anxious. The Associated Press (AP, 1987) reported Faigel's finding that students with unusually high test anxiety performed poorly; after taking a drug used to treat high blood pressure, their verbal scores rose by 50 points and their math scores by 70.

8. A weakness of this analysis is the lack of information about whether these students lived with their fathers, mothers, or both.

9. Seven of these items were easy (more than 80 percent of all test takers got them right). High r's on easy items are difficult to achieve, partly because errors unrelated to content—sloppy marks, using the wrong answer column, and the like—become an appreciable proportion of all errors, and such random errors act like "noise in the system" to reduce r's. On the 3 other items on which girls excelled, r = .37, modestly lower than the r on the items favoring boys.

10. The preceding analysis was originally prepared by James W. Loewen for inclusion in "Gender Bias in SAT Items."

11. Rosser is conducting additional research for the Commission on Testing and Public Policy to determine whether white and African American males and females use different problem-solving styles on SAT math questions.

12. This literature review and the complete bibliography at the conclusion of this report were prepared by Jacqueline Stevens. Works included were compiled from searches of three databases (ERIC, Psychological Abstracts, and ARLINE); also included are works discovered during the research for this report which are not included in these databases.

13. Sharif *et. al.* v. New York State Education Department *et. al.* 88 Civ. 8435. The case was brought as a class action suit by the Women's Rights Project of the American Civil Liberties Union (ACLU) on behalf of the Girls Clubs of America, the National Organization for Women, and 10 New York high school girls.

CHAPTER THREE

Closing Doors: The Impact of Sex-Biased Tests on Women's Educational Opportunities

Using SAT Scores to Award Merit Scholarships

Sex Bias in National Merit Scholarship Awards: Over $23 million in National Merit Scholarship awards, provided by 670 corporations, foundations, professional organizations, colleges and universities, are given annually to students with the highest scores on the Preliminary Scholastic Aptitude Test (PSAT). In 1987-88, women's average PSAT scores were 54 points lower than men's; women therefore were only 36 percent of the National Merit Scholarship semifinalists while 60 percent of the semifinalists were men (some students' gender could not be determined by their names). In 1986-87, 34.7 percent of the semifinalists were women (Rosser, 1987).

The semifinalist pool from which National Merit finalists and scholarship winners are chosen is based solely on the results of the PSAT, a test published by ETS and administered to high school juniors each October. PSAT score averages exhibit gender gaps similar to those on the SAT (13 points in the Verbal Section, 41 points in the Math—in SAT terms—in 1987-88). To qualify for the National Merit Scholarship, verbal scores are doubled and the math is added, in order to "give girls a better chance." But, as girls' verbal scores decline, doubling the verbal score is not overcoming the large gender gap in math scores. Students' PSAT scores must also be replicated by SAT scores in order for them to qualify as National Merit Finalists, so the bias on both these tests means that less scholarship money is awarded to girls. Talented young women also lose the prestige conferred on scholarship Semifinalists and Finalists that enhances college acceptance.

Hundreds of other merit scholarships are awarded annually to high school seniors by foundations, government agencies, unions, fraternal organizations, religious institutions, corporations (mainly sponsoring children of employees), professional organizations, and the military; these also use SAT scores to determine winners, either exclusively or in combination with grades. Most of these organizations refuse to provide a gender or racial breakdown of scholarship recipients so it is impossible to know the total amount of scholarship dollars girls are losing from these organizations.

■ **Using SAT Scores to Award State Merit Scholarships: State-by-State Analysis:** Almost half (22) of the States offer merit scholarships to high school seniors who choose to attend colleges or universities in their home state. A state-by-state survey of the 1988 awards was conducted as part of this study to see whether girls were receiving a fair share.

Massachusetts was the only state that relied solely on SAT scores to

determine scholarship winners, awarding scholarships to the top four SAT scorers in each district; the state scholarship office refused to give any male/female or ethnic data on the winners of the 160 scholarships awarded. *Missouri* uses only SAT or ACT scores; in 1988, winners needed to have a 680 SAT-Verbal and a 730 SAT-Math score or a score of 29 on the ACT (female/male data were not available on these awards).

Two other states—*Virginia* and *South Dakota*—use PSAT scores, awarding scholarships to all National Merit Semifinalists. Virginia also uses SAT and ACT scores; in 1988, *Virginia* nominated 50 students—29 women and 21 men, of whom only 10 were students of color. *South Dakota* awarded scholarships to 30 women and 38 men.

Eight other states use a combination of SAT scores and Grade Point Averages. Their 1988 awards for each sex are shown below:

State	Females	Males
Delaware	63	137
Florida	2,185	2,171
Georgia	310	340
Maryland	696	504
New Jersey	1,164	1,164
New York	12,325	12,575 (Regents)
	370	630 (Empire State)
Rhode Island	41	34
Tennessee	4	10

Six of the remaining states (*Arkansas, Idaho, Iowa, Louisiana, North Dakota,* and *Ohio*) rely solely on High School Grade Point Average, Class Rank and ACT scores. Four others (*Colorado, Illinois, Indiana, New Hampshire*) use only GPA and Class Rank. See page 107 for the complete State-by-State Survey findings.

In states where SAT scores are used in combination with grades and class rank, or are not used at all, girls generally receive more scholarships than boys. In states where SAT or ACT scores are used exclusively, boys are more likely to receive scholarships.

The outstanding exception is *New York*, which awards the most state merit scholarship money of any state—$8.24 million annually. In 1988, the New York State Department of Education decided to use a 50/50 formula of SAT (or ACT) scores and high school Grade Point Averages after years of relying solely on the SAT, because it was brought to the attention of the State Legislature that the SAT gender gap was preventing girls from receiving their fair share of these awards. Because of confusion in the reporting of grades in the first year, girls did not fare as well as expected, receiving only 37 percent (compared to 28 percent in the preceding year) of the 1000 Empire State Scholarships of Excellence ($2,000 per year for 5 years) and 50 percent of the Regents Scholarships ($250 per year for 5 years), even though girls were 53 percent of the test takers. For a variety of reasons, the State Department of Education decided to return to the exclusive use of the SAT in 1989.

In response, the Women's Rights Project of the American Civil Liberties Union brought suit against the State Education Department on behalf of the Girls Clubs of America, the New York chapter of the National Organization for Women, and 10 New York high school girls with grade point averages above 90. The suit, seeking to prevent New York from using the SAT as the

sole determinant of the awards, charged that women receive unequal consideration because they tend to score an average of 60 points lower than men on the SAT while consistently earning higher grades in New York's high schools. Since the purpose of the Empire State and Regents scholarships is to reward outstanding high school performance, not to predict first year college grades—the avowed purpose of the SAT—it was suggested that GPA rather than SAT scores be the selection criterion.

Although the State Education Department acknowledged that the SAT was not a perfect indicator of high school performance, it maintained as well that grades cannot be compared among schools because of grade inflation and because the collection process is too time consuming (U.S. District Court, Hearing Transcript, 1989). ETS filed an *amicus* brief on behalf of the Education Department, stating that using SAT scores for competitive scholarship awards is a "proper use" of the test, even though ETS has never indicated that the SAT evaluates high school performance, only that it predicts first year college performance (ACLU, 1988).

Federal District Judge John M. Walker did not agree, ruling instead that the use of SAT scores as the sole basis for awarding merit scholarships discriminates against girls (see Appendix I for the complete Opinion and Order). The State Education Department is developing its own scholarship exam and has requested funds from the State Legislature to complete the development and field test the new instrument in the fall of 1989 (U.S. District Court, Hearing Transcript, 1989).

■ **The Spin-off Effect:** Winners of State Merit Scholarships and National Merit Scholarships receive dozens of letters offering "no-need" scholarship awards, used by many colleges and universities to recruit high scoring students to attend their institutions. According to a 1984 study, more than 85 percent of the private four-year colleges and nearly 90 percent of the public institutions offer no-need scholarships for academic excellence (The College Board, 1984). In 1986, for example, a New Jersey senior who received a Garden State Distinguished Scholarship was also offered scholarships from 13 New Jersey colleges, 2 out-of-state colleges and 16 universities; eight other universities told him he qualified for their Honors Programs.

This spin-off effect is impossible to assess because it varies from student to student and state to state. However, it is important that parents and educators become aware of the interwoven nature of scholarship awards, in order to understand and appreciate the full extent of the financial and psychological damage inflicted by tests that do not predict classroom performance but do ensure access to important academic opportunities.

Using SAT Scores to Choose "Gifted and Talented" Students: State-by-State Analysis

Many states offer publicly-funded academic enrichment programs during the summer to high school students with high grades and high SAT, PSAT, or ACT scores (these do not include arts programs where admission is based on auditions or portfolios). A State-by-State Survey was conducted as part of this study to determine whether girls' educational opportunities at the middle and high school level were being limited by the use of these tests to select participants (see page 94 for complete State-by-State Survey findings).

Seventeen states use SAT, PSAT or ACT scores as part of their admissions formula. However, these test scores generally were used as 20 to 30 percent of an evaluation portfolio that included grades, essays, teacher recommendations, extra-curricular activities and demonstrated interest in the subject. Test scores therefore do not have an adverse effect on girls' participation in these summer programs; more girls than boys attend these programs, but involvement by both boys and girls of color is fairly limited. In fact, the evaluation process used by many states provided impressive alternatives to the exclusive use of college admission test scores.

States Using Test Scores	Females	Males
Alabama	45	35
Arkansas	200	200
Georgia	300	300
Hawaii	100	100
Iowa	30	28
Kentucky	327	314
Louisiana	222	180
Maine	32	28
Maryland	1,621	1,015
Mississippi	83	75
Missouri	155	165
New Jersey	135	165
Pennsylvania	132	158
Rhode Island	32	33
Tennessee	480	320
Texas	75	75
Virginia	Data not available	

■ **Private "Gifted and Talented" Programs—Exclusive Reliance on SAT Scores and Its Impact on Girls:** In contrast to these state programs, privately-funded summer programs for academically-talented 8th through

12th graders are far less open to girls. In the ten years since Johns Hopkins University began identifying "mathematically-precocious" children by administering the SAT to 7th graders, a number of similar talent search programs have been developed around the country. Academically-talented students are usually identified as those who score 430 or over on the SAT-Verbal and 500 or over on the SAT-Math as 7th graders; the score cut-off goes up 20 or 30 points for each grade above 7th. These students are then invited to attend a summer camp offering accelerated courses in math, science and the arts at the university sponsoring the talent search. For example, the Johns Hopkins program has now grown to five summer camps, held on both the east and the west coasts.

Six Talent Search programs were surveyed for this study, to assess the impact of girls' lower SAT score averages on their participation. It was not surprising that fewer girls participated in every progam that used SAT scores for admission:

Program	Criteria		Females	Males
Johns Hopkins Center for the Advancement of Academically Talented Youth (CTY)	up to 13-1/2 yrs. ≥ 430 SAT-V ≥ 500 SAT-M	Eligible students	15,162	14,879
		Qualified for CTY	2,584	3,316
		Attended Camp	1,181	1,556
Duke University Talent Identification Program	7th graders ≥ 500 SAT-V ≥ 550 SAT-M		339	528
		Ethnic Data		
		White	258	415
		Black	22	32
		Asian	50	60
		Hispanic	8	19
		Native Am	1	2
University of Denver Rocky Mountain Talent Search	7th graders ≥ 430 SAT-V ≥ 450 SAT-M for Humanities ≥ 500 SAT-M for Computer Science		40	80
University of California, Sacramento	ACT, SAT, SCAT above 90th percentile		197	246
		Cumulative Ethnic Data from all years		
		White	63%	
		Asian	16%	
		Black	3%	
		Hispanic	4%	
		Pilipino	1%	
		Other	5%	
University of California, Berkeley Academic Talent Development Program OSP (Older Student Program for 12-16 years)	12 year olds ≥ 440 SAT-V ≥ 460 SAT-M (20-30 points added for each year older)		250	287
		Ethnic Data		
		Asian	98	136
		White	83	88
		Black	35	21
		Hisp	16	22
		Native Am	1	2
		Other	17	18

One program surveyed—the ROGATE New Jersey Talent Search—used the high school achievement tests (CAT and ITBS) instead of SAT or PSAT scores and had 2,018 females and 1,835 males participating in the 1988 summer program. Since more males than females participated in all the other programs, it would seem that the use of SAT scores is keeping girls out of privately-sponsored summer programs for "gifted" students. Although it was impossible to determine the exact number of programs now operating in the country, it appears that an increasing number of girls are affected by these talent searches.

College Admissions—Are SAT Scores Essential?

The SAT or ACT is required for admission to nearly all of the 1500 four-year colleges and universities in the country (Rosser, 1987). Many colleges use strict cut-off scores; for example, the University of Texas at Austin requires that out-of-state applicants have a minimum combined SAT score of 1100 or ACT score of 27 (out of a possible 36) (Rosser, 1987).

Other colleges use test results in an admissions formula. The University of California at Berkeley, for example, adds the combined SAT score, the scores of three ETS Achievement Tests (where females also receive lower scores, except in writing and literature) and the Grade Point Average multiplied by 1000. In 1986, a combined number of over 7,000 was required for admission. Therefore, a candidate with a straight A (4.0) grade point average needed to score over 600 on all three achievement tests and over 1200 on the SAT.

Some universities require minimum test scores for admission to competitive programs. For example, Purdue University requires a 900 combined SAT score for admission to the engineering school (most students who are accepted have SAT combined scores of more than 1200); but Purdue has no SAT minimum for applicants in general. A number of universities require minimum SAT scores for entry into Honors programs, which often are similar to a small college within the university, with separate classrooms and residence halls.

Nearly every college in the country publishes average SAT scores for its incoming first year class and parents and high school guidance counselors use them to assist students in college selection. For example, Barrons' *Profiles of American Colleges* published the following score data on first year students entering several of the most selective colleges in 1988:

SAT—Verbal	Harvard	Yale	Stanford
Percent below 500	1%	2%	5%
Scored 500-599	15%	17%	20%
Scored 600-700	50%	46%	34%
Scored 700 or above	35%	34%	23%
SAT—Math			
Percent below 500	none	1%	1%
Scored 500-599	15%	9%	1%
Scored 600-700	50%	38%	35%
Scored 700 or above	35%	52%	56%

Minimum SAT score requirements dramatically affect African Americans, whose score averages in 1988 were 724 for females and 756 for males. Crouse and Trusheim note in *The Case Against the SAT* that a cut-off score of 900 would exclude 80 percent of the African Americans who took the SAT in 1984 and a cut-off of 1000 would exclude 90 percent (p. 95). A recent, controversial decision by the National Collegiate Athletic Association (NCAA) brought this issue to light in a different context.

The NCAA proposed to require student athletes to have a GPA of 2.0 and a minimum SAT score of 700 or ACT of 15 in order to receive a college athletic scholarship (Proposition 42). This is a restriction that goes beyond the earlier Proposition 48 to exaggerate the importance of SAT scores; Proposition 48, which went into effect in 1985, required incoming first year athletes to have a 2.0 GPA in a core curriculum and a minimum SAT score of 700 or ACT of 15. Those who satisfied only one requirement were permitted to receive scholarships but were not permitted to play or practice during their first college year.

Proposition 42 would amend Proposition 48 to require students to meet *both* requirements in order to receive scholarships and, in most cases, to attend college; it will have an adverse impact on many low income students of color, who have the greatest need for scholarship aid.

The NCAA stated that the purpose of this change was to tighten entrance requirements so that athletes who play sports for a college also will be able to graduate from their school. But Georgetown University basketball coach John Thompson challenged this reasoning. He opposed imposition of Proposition 42, stating that standardized tests were culturally-biased against students of color and low income students, did not sufficiently predict their academic ability, and should not be used to restrict their access to scholarships (New York *Times*, January 21, 1989).

Loewen, Rosser, and Katzman (1988) conducted a study of 1,112 New York City high school students and found that some females with A+ GPAs but lower SAT scores had self-selected themselves out of the elite college pool. They were not planning to apply to the most competitive colleges at the same rate as boys with similar grades. In fact, girls in all 4 GPA areas studied planned to go to slightly less prestigious colleges than boys with equivalent GPAs.

College admissions officers often use a mathematical formula that combines high school grades and SAT scores, weighting them in a way that predicts how well students are supposed to do in their first college year. If the same equation is used for both sexes, girls are predicted to do less well in college than they *actually* do (by one-fourth to a full standard deviation below their actual GPA), according to a 1973 study by the American College

Testing Program. A separate equation for girls more closely predicts college performance. But, if this were the determining factor, according to Nancy Cole, former president of the National Council of Measurement in Education, "they'd be accepting two-thirds girls to one-third boys. Since this isn't happening, we know some other criteria are involved." Although those criteria are not public information, Cole suggested (in telephone coversation with Rosser in 1979) that quotas are used.

In fact, the advantage the SAT gives men in admissions may be one of the reasons that some universities and colleges rely on it. In 1987, the Washington *Post* reported that the University of North Carolina at Chapel Hill had downplayed its use of the SAT with the result that the university's enrollment was nearly two-thirds female. One trustee voiced strong opposition, suggesting that the SAT be reinstated so more men would be admitted. He was concerned, he said, that the University would lose political clout in the State House and alumni dollars, since women give less money to their alma maters.

■ **Princeton University—A Case Study of Underprediction:** Even women who are accepted at the most competitive universities find their SAT scores underpredicting their college performance. In an unpublished senior thesis for Princeton University, Julie Lubetkin compared the grades, courses and SAT scores of the Princeton University Class of 1990, who entered college in the fall of 1986. The women's average SAT scores were slightly higher in the Verbal Section but considerably lower in the Math, although not as low as the national average. Even though their average SAT scores were lower than the men's, their average first year GPA was slightly higher. In other words, SAT scores underpredicted the women's grades and overpredicted the men's grades, with the SAT-Math being the significant underpredictor. Lubetkin found that in at least half of the academic departments at Princeton, one section of the SAT was not helpful in predicting students' grades. Although the differences were not large, it is likely that such an underprediction of men's grades at Princeton would not be tolerated by the university. Lubetkin also refers to a 1985 study by Strenta of the Dartmouth first year class, which found the SAT less predictive for women, who earned higher first year grades than men but lower combined SAT scores.

■ **The Massachusetts Institute of Technology, Bates College, and Bowdoin College—New Admissions Policies To Counter the SAT's Underprediction for Women:** Some universities have taken action against the SAT's underprediction of women's academic performance. The Massachusetts Institute of Technology's Admissions Office conducted a study of student performance and discovered that females with lower SAT-Math scores were achieving Grade Point Averages equal to or better than their male peers in their sophomore and senior years. Michael Behnke, Director of Admissions, says that this study "excluded the possibility that it is due to differences in course selection by men and women. Women also have a higher retention rate so it is not due to women dropping out at a higher rate." As a result, MIT has been admitting women with lower SAT scores than men (Behnke, 1987).

Several other colleges have dropped the use of the SAT altogether, including Bates and Bowdoin in Maine, Middlebury College in Vermont and Union College in Schenectady, New York. A two-year study at Bates, according to William Hiss, found that the applicant pool increased 17.6 percent, with significant increases in geographical diversity, minority applications and foreign applications (Crouse and Trusheim, 1988).

Applicants who chose not to submit SAT scores averaged 80 points lower on both the SAT Verbal and Math Sections than applicants who submitted their scores, but they did not differ significantly in first year GPA or academic standing.

William A. Mason III, Director of Admissions at Bowdoin, reports that SAT scores were made optional in 1969 "in order to encourage minorities and the economically disadvantaged to apply." He says that Bowdoin evaluates applications by asking each high school to rank their academic courses in order of difficulty and then looking to see if applicants took the "full program of courses." Only after the quality of courses has been evaluated do admissions staff look at Grade Point Average. Three thousand seven hundred people applied for the 385 openings for the class of 1991, and a third of these applicants did not provide SAT scores. But the admissions officers were able to make difficult admissions decisions "relying minimally, if at all, on the Educational Testing Service exams. In a climate where parents, guidance counselors and school boards all overemphasize the importance of test scores, we believe that our process is the fairest" (ACLU, 1988).

1988 Demographic and Selection Data for State-Sponsored Summer Programs for Gifted High School Students[1]

State:	**ALABAMA,** Governor's School.	
Total Students:	80	
Female:	45	
Male:	35	
Ethnic Data:		
White:	74	
Black:	F-2	M-1
Hispanic:	F-1	
Asian American:	F-1	
Other:		M-1
Criteria:	SAT, PSAT, ACT, I.Q., G.P.A., Essay, Other.	

Selection Formula: *Count 75%*: SAT-1100, PSAT-170, ACT-25 OR I.Q. 120+ on Leiter, Ravens, S-B, WISC-R. *Count 25%*: G.P.A. 2 Teacher recommendations, Principal or counselor recommend. Essay.

Comments: Program partially state supported. Funding must be renewed every year.

State: **ALASKA,** No Program.

State:	**ARIZONA,** Northern Arizona University Career Institute (for minorities).
Total Students:	150
Female:	75
Male:	75

Selection Formula: Counselor, Teacher, Community recommendations only.

1. Compiled by Phyllis Rosser and Jacqueline Stevens.

State:	**ARKANSAS,** 1988 Arkansas Governor's Program (GP). 1987 Academic Enrichment for Gifted in Summer (AEGIS).	
Total Students:	GP 400	AEGIS 1030
Female:	200	576
Male:	200	463
Ethnic Data:		
White:	346	909
Black:	37	121
Hispanic and Asian American:	17	
Criteria:	SAT, PSAT, ACT, Achievement, Other.	

Selection Formula: SAT, PSAT, ACT 90th percentile. Exhibit exceptional ability but not high scores. Grades in all courses. Student essay. School recommendations. Arts—require portfolios/auditions; no tests.

Comments: State selection committee choose Governor's school. Directors of AEGIS select their students.

State: **CALIFORNIA,** No Program.

State: **COLORADO,** No Program.

State: **CONNECTICUT,** Center for Creative Youth: Visual and Performing Arts/Creative Writing.

Selection Formula: Portfolios/Auditions/Writing samples.

Comments: Mostly District funded, some private funds.

State:	**DELAWARE,** Governor's School for Excellence.	
Total Students:	114	
Female:	64	
Male:	50	
Ethnic Data:		
White:	F-45	M-32
Black:	7	4
Hispanic:	1	0
Asian American:	7	9
Other:	5	5 (Native Americans)
Criteria:	Essay	

Selection Formula: H.S. Discretion. One essay required. Arts-require portfolios/auditions; no tests.

Comments: One student chosen for every 600 in H.S.

State:	**DISTRICT OF COLUMBIA,** No Program.

State:	**FLORIDA,** Governor's Summer Program.
Total Students:	171
Female:	109
Male:	62
Ethnic Data:	
White:	148
Black:	13
Hispanic:	4
Asian American:	6
Criteria:	Identified gifted. GPA. Test Scores.

Selection Formula: Previously identified as gifted (2 standard deviations above the mean on Standardized I.Q. test) or Demonstrated High Achievement (any measure may be used such as GPA, St. Achievement tests).

Comments: Both the gifted and high achieving students are eligible so that any highly motivated student can attend.

State:	**GEORGIA,** Governor's Honors Programs in eleven instructional areas: English, Science, Social Studies, Math, Foreign Language, Visual Arts, Music, Dance, Theatre, Entrepreneurship, Design.
Total Students:	600
Female:	300
Male:	300
Criteria:	PSAT, GPA, CTBS/ITBS, Other.

Selection Formula: Math & Science-programs require PSAT. Arts—Portfolio/Audition. Others require combination of Grades/CTBS or ITBS. Students' written statement. Teacher recommendations citing evidence of high interest in subject.

Comments: Use PSAT because it's broader, less of an achievement test in verbal and computational skills than SAT.

State:	**HAWAII,** Summer Program (SP) and Enrichment in Language Arts (ELA).	
	SP	ELA
Total Students:	200	240
Female:	100	120
Male:	100	120
Criteria:	SAT, PSAT, Grades, Other.	

Selection Formula: SAT, PSAT-Count 50%, Grades-Count 25%, Extracurricular Activities-Count 25%.

State: **IDAHO,** No Program.

State: **ILLINOIS,** No Program.

State: **INDIANA,** Starting Governor's Scholars Academy Summer 1989.

Selection Formula: Plan to use: Test Scores, Equal Numbers for each county, plan to have 15% of Students from "under-represented" populations.

State: **IOWA,** Governor's Institute For Gifted and Talented: Science, Math, Humanities.

Total Students:	Science	Math	Humanities
	24	21	23
Female:	10	8	12
Male:	10	10	8
Ethnic Data:			
Minorities:	4	3	3
Criteria:	SAT, PSAT, ITBS, ITED, Other.		

Selection Formula: SAT, PSAT, Achievement tests: ITBS, ITED. Special reading and math test to place skills by grade levels. Nomination by school district.

State: **KANSAS,** Kansas Regents Honors Academy (1987 figures).

Total Students:	128	
Female:	62	
Male:	66	
Ethnic Data:		
White:	F-57	M-58
Black:	2	1
Hispanic:	3	1
Asian American:	0	4
Other:	4	
Criteria:	Any Standardized Test.	

Selection Formula: Standard tests-count 10%, Grades 10%, Extracurricular Activities 30%, Class Rank 10%, Teacher Recommends (2) 10%, Teacher Checklist (2) 6%, Student Essay 15%, Senatorial District Ranking 4%, Evaluator Ranking 5%.

State:	**KENTUCKY,** Governor's Scholars Programs apportioned by county for rural/urban balance: W. Kent (rural)-21%, N. Kent (urban)-10%, Jefferson (urban)-16%, Cent. Kent-22.9%, Appalachia (rural) 28%.	
Total Students:	641	
Female:	327	
Male:	314	
Ethnic Data:		
White:	F-288	M-294
Black:	27	11
Hispanic:	0	0
Asian American:	12	9
Criteria:	PSAT, H.S. Achieve. Tests, Essay, Other.	

Selection Formula: PSAT (verbal looked at more closely than math) St. achievement test-96 percentile. Ranked on Renzulli scale. Teacher recommendations (2+) to indicate special interest or talent. Essay: experiences/community life most influential.

Comments: Don't want students that may be a bit rigid or that have the best grades. Local selection committee includes 2 community members. Final selection made by state committee.

State:	**LOUISIANA,** State School of Math, Science and the Arts. Residential-August to May. 1987 data.	
Total Students:	402	
Female:	222	
Male:	180	
Ethnic Data:		
White:	F-173	M-147
Black:	18	10
Hispanic:	0	0
Asian American:	27	22
Other:	4	1
Criteria:	SAT, CTBS, SRA, GPA, Interview, Other.	

Selection Formula: Use 17 Criteria which include: GPA+SAT+CTBS/SRA. Recommendation by principal or guidance counselor re: leadership, commitment, inquisitiveness, potential for success. Recommendations by teacher, employers, church. Interview/audition at school. Faculty interviews.

Comments: Reviewed by State Selection Committee.

State:	**MAINE,** Maine Summer Humanities Institute.
Total Students:	60
Female:	32
Male:	28
Criteria:	PSAT, Essay, Other.

Selection Formula: PSAT (not heavily weighted). Essay Questions (most weight). Teacher Recommendations, Out-of-School Educational Experiences.

Comments: Also offers a Summer Arts Program.

State:	**MARYLAND,** Grades 4-12, 11 Summer Institutes: Aquatic Studies, Critical and Creative Thinking, Science Internships in Government and Business, Leadership, Space Science (at Goddard), Chesapeake Bay Studies, COMPETE (Math), International Studies, Archeological Research, The Lady Maryland Experience (Environmental Studies aboard a schooner).
Total Students:	2636
Female:	1621
Male:	1015
Ethnic Data:	
Minorities:	580
Criteria:	SAT/PSAT, H.S. Achievement Tests, Essay, Other.

Selection Formula: SAT, PSAT count-33.3%, Student Essay-33.3%, Teachers Recommendations-33.3%, No grades but standardized Achievement Test Stanines are looked at.

Comments: Allocated to each county.

State:	**MASSACHUSETTS,** School/College Collaborative Programs (held at 10 colleges). Ronald E. McNair Programs (for educationally disadvantaged).
Total Students:	4831 (Collab) 548 (McNair)
Criteria:	Varies with College

Selection Formula: Collaborative Programs: varies with each college. Grades, extracurricular activities, teacher recommendations used more frequently than test scores. McNair Programs: family income, race, school record.

State:	**MICHIGAN,** State Board of Education, Summer Institutes for the Arts & Science.
Total Students:	540
Female:	323
Male:	212
Ethnic Data:	
White:	406
Black:	80
Hispanic:	14
Asian American:	26
Other:	14

Selection Formula: Look for students who: show unusual interest, ability, involvement; show motivation, commitment, direction, curiosity; are creative or exceptionally talented; have the potential but haven't had a chance to pursue it in an intense environment.

Comments: Student does not have to be in gifte !/talented program to apply.

State: **MINNESOTA,** No Program.

State:	**MISSISSIPPI,** Governors School (partially funded by State legislature).
Total Students:	158
Female:	83
Male:	75
Criteria:	SAT, PSAT, ACT, Essay, Other.

Selection Formula: *Weigh heaviest (1.2.3)* 1. SAT-1100 minimum score, 2. PSAT-175 (NMQT score), 3. ACT-25 minimum score. 125 or above on Leiter Interest Performance Scale, Raven's Standards, Stanford-Binet, WISC-R or WAIS.

Comments: Nominated by H.S.

State: **MISSOURI,** Missouri Scholars Academy. (20-50 places are allotted to create minority, geographic, and gender balance).

Total Students:	320	
Female:	155	
Male:	165	
Ethnic Data:		
White:	F-116	M-130
Black:	17	14
Hispanic:	0	2
Asian American:	7	8
Other:	15	11
Criteria:	PSAT, WISC, GPA, Other.	

Selection Formula: PSAT-20%, WISC or WAIS-20%, GPA-10%, 2 Essays plus evidence of outstanding abilities-50%.

Comments: Each H.S. submits at least one student. Large districts submit top 1% of class. Look for gender, racial, geographical balance.

State: **MONTANA,** No Program.

State: **NEBRASKA,** No Program.

State: **NEVADA,** Governor's Summer Institute.

Total Students:	60
Female:	41
Male:	19
Criteria:	Outstanding academic or creative performance.

Selection Formula: H.S. discretion based on: good academics, creative abilities, leadership, some H.S. require essay.

Comments: One student chosen from each H.S.

State: **NEW HAMPSHIRE,** No Program.

State:	**NEW JERSEY,** Governor's School: Science, Public Issues, The Arts.		
	Science	Issues	Arts
Total Students:	100	100	100
Female:	35	50	50
Male:	65	50	50
Ethnic Data:			
White:	F-26 M-41	F-32 M-32	F-50 M-40
Black:	3 2	9 3	4 3
Hispanic:	2 3	3 7	3 3
Asian American:	4 19	6 8	6 4
Criteria:	SAT, PSAT, ACT, GPA, Other.		

Selection Formula: Science: PSAT-30%, Grades-20%, Extracurricular activities-15%, Recommendations from teachers-20%, 2 Essays-15%. Public Issues: SAT/PSAT/ACT-90th percentile or above-count 5%, Grades-B average or above-5%, Extracurricular activities- 20%, Recommendations-70%, Schools are asked to look for 19 characteristics of gifted students like risk taker, good guesser, etc. The Arts: Audition/Portfolio.

Comments: Science: Each school nominates one student. Larger schools nominate more. Public Issues: Although tests only count 5% they have more impact because of minimum requirement. Chosen by school, county, outside evaluator and college where school is held.

State: **NEW YORK,** Six Summer Schools for the Arts: Choral Studies, Orchestra Studies, Theatre, Film/Media, Dance, Visual Arts. To begin in summer 1989: Math and Science Summer Institutes.

Selection Formula: The Arts: Auditions and Portfolios. Math and Science Institute: Criteria not yet determined.

Comments: SAT not to be used for Science.

State: **NEW MEXICO,** No Program.

State:	**NORTH CAROLINA,** Governor's School.	
Total Students:	807	
Female:	428	
Male:	379	
Ethnic Data:		
White:	F-332	M-310
Minorities:	96	69
Criteria:	GPA, Achievement Tests, I.Q. Tests.	

Selection Formula: Use point system based on grades, achievement tests, I.Q. tests and other standardized tests: Weight for each by H.S. discretion.

State: **NORTH DAKOTA,** No Program.

State: **OHIO,** 13 Summer Institutes in Humanities, Arts, Science, Engineering held at State Universities. Bryl R. Shoemaker School for Vocationally-talented students at Denison University. Martin W. Essex School for the Gifted at Ohio State University.

Total Students: 2000

Criteria: GPA, Essay, Other. Bryl R. Shoemaker: Nomination by superintendent. Martin W. Essex School: Essay, Other.

Selection Formula: H.S. discretion varies by program but usually includes: demonstrated talent or interest, grades, essay, teacher recommendations. Some require 89th percentile on standardized I.Q. or Achievement test. Martin W. Essex School: H.S. discretion, essay. 2 teacher recommendations, Demonstrated creativity, leadership, etc.

Comments: Each district chooses one student or one for every 10,000 students. Martin W. Essex School: Selected by State Committee.

State: **OKLAHOMA,** No Program.

State: **OREGON,** No Program.

State: **PENNSYLVANIA,** Governor's School: Science, Business, International Affairs, Agriculture, Arts.

Total Students:	S:	B:	IA:	Ag:
	98	65	62	65
Female:	39	30	30	33
Male:	59	35	32	32

Criteria: SAT/ACT, H.S. Rank, Essay, Other.

Selection Formula: Science: ACT/SAT (most are between 1400-1600). Score on Westinghouse test for science. H.S. rank, Teacher recommend. Essay (heaviest weight). Business, Inter. Affairs, Agriculture look at: SAT/ACT scores, H.S. rank. Essay-to indicate interest. Teacher recommendations on ability. Arts: portfolios and auditions.

Comments: Students must be interested in subject and have demonstrated ability. A subjective decision is made.

State:	**RHODE ISLAND,** Governor's Summer Program in Science and Mathematics.	
Total Students:	65	
Female:	32	
Male:	33	
Ethnic Data:		
White:	F-29	M-30
Asian American:	3	1
Other:		2 (Native Americans)
Criteria:	SAT/ACT, H.S. GPA, Other.	

Selection Formula: SAT, ACT-20%, GPA-60%, Teacher or counselor recommendations-20%.

Comments: Nominated by teachers or guidance counselors.

State:	**SOUTH CAROLINA,** Governor's School.	
Total Students:	238	
Female:	128	
Male:	110	
Ethnic Data:		
White:	F-95	M-80
Black:	31	19
Hispanic:	0	1
Asian American:	2	10
Criteria:	H.S. Rank, SAT/PSAT/BSAP, Essay, Other.	

Selection Formula: SAT/PSAT/BSAP-carry small weight. Class Rank-top 5%. *Carry most weight.* GPA, Essays, Teacher Recommendations, Extracurricular activities.

Comments: A balance is sought in gender, race, and geographical representation.

State:	**SOUTH DAKOTA,** Governor's Camp for the Gifted.	
Total Students:	150	
Female:	75	
Male:	75	
Ethnic Data:		
White:	128	
Other:	22 (Native Americans)	
Criteria:	Stanford Achievement Test.	

Selection Formula: Must score in 98th percentile on Stanford Achievement Test. Selection made on basis of gender and zip code.

State:	**TENNESSEE,** Governor's Schools for: Arts, Humanities, Science, Math, International Studies, Tennessee Studies.	
Total Students:	800	
Female:	480	
Male:	320	
Ethnic Data:		
White:	F-420	M-258
Black:	26	24
Hispanic:	12	12
Asian American:	22	26
Criteria:	SAT/PSAT/ACT, H.S. GPA, Other.	

Selection Formula: SAT, PSAT, ACT-20%; H.S. GPA-20%; Extracurricular activities-20%; Teacher Recommendations-20%; Student work sample-20%.

State:	**TEXAS,** Governor's School.	
Total Students:	150	
Female:	75	
Male:	75	
Ethnic Data:		
White:	F-38	M-41
Black:	10	5
Hispanic:	15	13
Asian American:	11	16
Criteria:	PSAT, Achievement Tests, H.S. GPA.	

Selection Formula: Use matrix of 7 items, all of which have equal weight and the lowest criterion is dropped. Matrix includes: PSAT, Achievement Tests, H.S. GPA.

State: **UTAH,** No Program.

State:	**VERMONT,** Governor's School in: Science and Technology. International Affairs, The Arts.	
	ST	IA
Total Students:	66	62
Female:	25	37
Male:	41	25

Selection Formula: No test scores used. 59 supervisory unions select students based on own criteria. Look for potential, not proven performance; those who want explorational, creative, inspirational experience rather than disciplined training.

State:	**VIRGINIA,** Governor's Schools (Residential): Humanities, Mathematics, Science & Technology, Japanese Language & Culture; NASA/Virginia Institute of Marine Science Mentorship; Visual and Performing Arts; Foreign Language Academics.
Total Students:	Humanities: 200
	Mentorship: 44
	Visual/Performing Arts: 130
	Foreign Language: 130
Criteria:	SAT/PSAT/STEA/EAS, Standarized Achievement Tests. GPA, Essays.

Selection Formula: Honors Received, 2 Essays, 2 teacher recommendations. SAT/PSAT/STEA/EAS, Standardized Achievement Test scores as SRA comp, CAT, NEDT, Stanford Ach, ITBS. GPA and difficulty of courses taken. Visual and Performing Arts: Audition or work review. Foreign Language: must demonstrate proficiency through tapes, written composition. Teachers' recommendations also important.

State: **WASHINGTON,** No Program.

State:	**WEST VIRGINIA,** Governor's Honor Academy.
Total Students:	147
Females:	81
Males:	66
Ethnic Data:	
Whites:	138
Blacks:	2
Hispanics:	0
Asian/Americans:	6
Other:	1 (Native American)
Criteria:	CTBS

State: **WISCONSIN,** No Program.

State:	**WYOMING,** U.W. Summer High School Institute.
Total Students:	91
Females:	52
Males:	38
Ethnic Data:	
Whites:	86
Blacks:	0
Hispanics:	0
Asian Americans:	4
Other:	1 (Native American)

Selection Formula: *Counts 50%:* Letter of application describing what they think they'll bring to the program and get out of it. One page essay-about a project, interest, or activity that represents them as a person. *Counts 25%:* 2 Teacher recommendations following a format to distinguish unusual qualities. *Counts 25%:* Extracurricular activities, Standardized test scores, GPA-counts least because gifted students may not have high grades due to boredom, etc.

Comments: Every H.S. is assured one student accepted. No attention is paid to gender balance. Only 10th graders are involved, to maximize the benefits of the program in their high school years. Courses are for broadening student's perspective rather than acceleration.

[redacted]

1988 Survey of Statewide Merit Scholarship Programs[1]

State:	**ARKANSAS**
Program:	Governor's Scholarship Program.
Students:	345
Average Award 1987-88:	2,000
Total Payout (Millions):	0.69
Criteria:	ACT, H.S. GPA, other.
Formula:	ACT cutoff score (26) + unweighted GPA (3.6 minimum) + rank in class + leadership.
Comments:	Students' activities on application reviewed by State panel.
Females:	199
Males:	146
Ethnic Data:	329 White (95.36%), 13 Black (3.77%), 2 Asian American (.58%), 1 Hispanic (.29%).

1. Compiled by New York Public Interest Group (NYPIRG) and Phyllis Rosser.

State:	COLORADO
Program:	Undergraduate Merit Awards.
Students:	10,700
Average Award 1987-88:	Varies by year.
Total Payout (Millions):	7.0
Criteria:	H.S. GPA, College GPA.
Formula:	Colleges divide money allotted by state; normally ranked by GPA.
Comments:	No state guidelines; college discretion.
Females:	N.A.
Males:	N.A.
Ethnic Data:	N.A.

State:	DELAWARE
Program:	Diamond State Scholarships.
Students:	200
Average Award 1987-88:	1,000
Total Payout (Millions):	0.2
Criteria:	SAT, H.S. GPA, H.S. Rank, Other.
Formula:	SAT + H.S. GPA + H.S. Rank + AP courses + Guidance counselor evaluation counts 88%, Activities in School/Community, special awards and essays count 12%.
Comments:	No Statewide GPA Guidelines; Guidance Counselors provide evaluation of courses and activities. State evaluates GPA and essays.
Females:	63
Males:	137
Ethnic Data:	N.A.

State:	**FLORIDA**
Program:	Undergraduate Scholars Fund.
Students:	4,626
Average Award 1987-88:	1,000 to 2,500
Total Payout (Millions):	3.85
Criteria:	SAT/PSAT/ACT.
Formula:	National Merit Finalists or 1200/28 SAT/ACT or unweighted 3.5 GPA or International Baccalaureate or Florida Academic Scholars Certificate (1100 SAT; 26 ACT; 3.0 GPA).
Comments:	Must meet one requirement for eligibility and funding. Award depends on resources available in program.
Females:	2,185
Males:	2,171 (270 sex unknown)
Ethnic Data:	N.A.

State:	**GEORGIA**	
Program:	Governor's Scholarship Program.	
Students:	650	
Average Award 1987-88:	1,275	
Total Payout (Millions):	0.83	
Criteria:	SAT, H.S. GPA, Other.	
Formula:	1300 SAT (one sitting) + 3.75 unweighted GPA, 3 sports, leadership outside school.	
Comments:	State assesses activities; everyone eligible receives award; H.S. discretion for GPA (no Guidelines).	
Females:	310	
Males:	340	
Ethnic Data:	Female	Male
	290 White	314 White
	6 Black	7 Black
	3 Hispanic	4 Hispanic
	9 Asian Am.	11 Asian Am.
	2 Unknown	4 Unknown

State:	IDAHO
Program:	State of Idaho Scholarships.
Students:	76 (includes vocational scholarships).
Average Award 1987-88:	1,500
Total Payout (Millions):	.114
Criteria:	ACT, H.S. Rank.
Formula:	Percentile rank is converted into a class rank "standard score".
Comments:	ACT/class rank score count 50/50 (Recent change from weighted GPA). Class rank score calculated with math formula.
Females:	37
Males:	39
Ethnic Data:	N.A.

State:	ILLINOIS
Program:	Merit Recognition Scholarships.
Students:	4,402 (Renewals only. No new awards in 1987-88 due to lack of funds).
Average Award 1987-88:	500
Total Payout (Millions):	2.2
Criteria:	H.S. GPA.
Formula:	Must be in top 10% H.S. Rank + unweighted GPA (top 10% = eligibility, State uses GPA to rank candidates.)
Comments:	H.S. discretion.

State:	INDIANA
Program:	Hoosier Scholarships.
Students:	815 (1988-89)
Average Award 1987-88:	500
Total Payout (Millions):	0.40
Criteria:	H.S. Rank, Other.
Formula:	Top 20% of class. H.S. recommendation based on courses/GPA.
Comments:	H.S. discretion; top 1-3 students from each High School.
Females:	482
Males:	333
Ethnic Data:	N.A.

State:	**IOWA**
Program:	Scholarship Program.
Students:	2,107
Average Award 1987-88:	200-600
Total Payout (Millions):	0.736
Criteria:	H.S. Rank, H.S. GPA, ACT.
Formula:	Top 15% H.S. Rank eligible; then ranked by formula using GPA + ACT.
Comments:	H.S. discretion (GPA normally unweighted) 50/50 formula.
Females:	1,072
Males:	940
Ethnic Data:	N.A.

State:	**LOUISIANA**
Program:	T.H. Harris Scholarships.
Students:	3,749
Average Award 1987-88:	200-300/year
Total Payout (Millions):	0.55
Criteria:	H.S. GPA, ACT.
Formula:	3.0 GPA = eligibility; weighted GPA used to rank students (ACT used as tie breaker).
Comments:	Honors weighted encouraged with H.S. discretion.
Females:	N.A.
Males:	N.A.
Ethnic Data:	3,234 White, 515 Black.

State:	**MARYLAND**
Program:	Distinguished Scholar Program.
Students:	1,200
Average Award 1987-88:	1,600
Total Payout (Millions):	1.859
Criteria:	H.S. GPA,SAT,PSAT,ACT.
Formula:	Unweighted GPA primary criteria. SAT, PSAT, ACT used as tie breaker.
Females:	696
Males:	504
Ethnic Data:	N.A.

State:	**MASSACHUSETTS**
Program:	Honor Scholarships.
Students:	160
Average Award 1987-88:	936-1296 (free tuition to 13 state colleges).
Total Payout (Millions):	0.750
Criteria:	SAT
Formula:	Awarded to top four SAT scores in each district.
Comments:	Awarded by district.
Females:	N.A.
Males:	N.A.
Ethnic Data:	N.A.

State:	**MASSACHUSETTS**
Program:	Commonwealth Scholars.
Students:	1,059 (1987)
Average Award 1987-88:	1,000
Total Payout (Millions):	.930
Criteria:	H.S. GPA.
Formula:	3.5 GPA + 2 teacher recommendations + extracurricular and community activities.
Comments:	H.S. discretion.
Females:	N.A.
Males:	N.A.
Ethnic Data:	N.A.

State:	**MISSOURI**
Program:	Higher Education Academic Scholarships.
Students:	1,378 (approved 1988), 860 (returning).
Average Award 1987-88:	2,000
Total Payout (Millions):	1.813
Criteria:	SAT/ACT
Formula:	Top 3% in SAT (1360 minimum score) score or ACT (29 minimum). 1988 winners needed: 730 Math, 680 Verbal.
Comments:	Must have both Math and Verbal minimums. Program in second year.
Females:	N.A.
Males:	N.A.
Ethnic Data:	N.A.

State:	NEW HAMPSHIRE
Program:	Governor's Scholars Awards.
Students:	181
Average Award 1987-88:	700
Total Payout (Millions):	0.126
Criteria:	H.S. Rank.
Formula:	Valedictorian Awarded, 2nd-4th if graduating class over 100.
Comments:	H.S. discretion.
Females:	104
Males:	77
Ethnic Data:	N.A.

State:	NEW JERSEY
Program:	Distinguished Scholars Program.
Students:	2,328 (1988)
Average Award 1987-88:	1,000
Total Payout (Millions):	2.328
Criteria:	H.S. Rank + SAT.
Formula:	1st, 2nd in school win: others in top 10% H.S. Rank + 1200 SAT win.
Comments:	H.S. discretion.
Females:	1,164
Males:	1,164
Ethnic Data:	N.A.

State:	NEW YORK
Program:	Regents College Scholarships.
Students:	25,000 (1988)
Average Award 1987-88:	250/yr. (renewed up to 5 years)
Total Payout (Millions):	6.2M
Criteria:	SAT + H.S. GPA
Formula:	SAT + GPA count 50/50.
Comments:	Awarded by County depending on no. of H.S. students from previous year.
Females:	12,325
Males:	12,575
Ethnic Data:	20,950 White, 950 Black, 700 Hispanic, 1,850 Asian Americans, 50 Native Americans, 300 Other, 2,000 Unknown.

State:	**NEW YORK**
Program:	Empire State Scholarships of Excellence.
Students:	1,000 (1988)
Average Award 1987-88:	2,000 (renewed up to 5 years.)
Total Payout (Millions):	2.0
Criteria:	SAT + H.S. GPA.
Formula:	SAT + GPA count 50/50.
Comments:	Awarded by County depending on no. of H.S. Students.
Females:	370
Males:	630
Ethnic Data:	793 White, 10 Black, 11 Hispanic, 162 Asian American, 3 Native American, 12 Other 9 Unknown.

State:	**NORTH DAKOTA**
Program:	Scholars Program.
Students:	48
Average Award 1987-88:	1,068-1,194
Total Payout (Millions):	.055
Criteria:	ACT, H.S. GPA.
Formula:	Must be in upper 5th percentile of ACT + upper 20th percentile of H.S. Rank to be eligible.
Females:	16
Males:	32
Ethnic Data:	N.A.

State:	**OHIO**
Program:	Academic Scholarship Program.
Students:	4,000
Average Award 1987-88:	1,000
Total Payout (Millions):	4.0
Criteria:	ACT, H.S. Rank, H.S. GPA.
Formula:	ACT + H.S. Rank (by percentile) + H.S. unweighted GPA.
Comments:	Index is constructed combining H.S. Rank + GPA (50%) and ACT (50%); H.S. discretion. Top 5 students in each H.S. usually receive awards.
Females:	2,188
Males:	1,812
Ethnic Data:	3,696 White, 96 Black, 32 Hispanic, 148 Asian American, 28 Other.

State:	RHODE ISLAND
Program:	Governor's Academic Scholarships.
Students:	75
Average Award 1987-88:	420 to 2,500
Total Payout (Millions):	.141
Criteria:	SAT, ETS Achievement tests, H.S. Rank, H.S. GPA.
Formula:	Matrix of H. S. Rank, GPA, SAT, top 2 Achievement tests.
Comments:	State colleges determine own winners. Criteria vary.
Females:	41
Males:	34
Ethnic Data:	N.A.

State:	SOUTH DAKOTA
Program:	Superior Scholar Scholarships.
Students:	68
Average Award 1987-88:	1,404
Total Payout (Millions):	.0955
Criteria:	PSAT
Formula:	Award National Merit Semifinalists only.
Females:	30
Males:	38
Ethnic Data:	N.A.

State:	TENNESSEE
Program:	Academic Scholars Program.
Students:	14
Average Awarded 1987-88:	4,000
Total Payout (Millions):	.056
Criteria:	SAT/ACT, H.S. GPA.
Formula:	SAT/ACT (65%) + GPA (27.5%) + Extracurricular activities (2.5%) + Upper division credits in math, science, foreign language (5%).
Females:	4
Males:	10
Ethnic Data:	N.A.

State:	**VIRGINIA**
Program:	Virginia Scholars Program.
Students:	50
Average Awarded 1987-88:	3,000
Total Payout (Millions):	.150
Criteria:	SAT/PSAT/ACT.
Formula:	Semifinalists in National Merit Scholarship and Achievement programs automatically nominated. Other scores requested but not mandated: ETS Achievement, AP, H.S. Achievement.
Comments:	H.S. nominates students. University admissions Directors pre-screen nominees for grades, academic honors, leadership, extracurricular activities. Winners chosen by State Awards Selection Committee.
Females:	29
Males:	21
Ethnic Data:	40 White; 10 Minority Students.

CHAPTER FOUR

Recommendations for Further Research and Development

The findings of this and other studies suggest that the test publisher could take certain steps immediately to address the SAT's underprediction and bias; these are summarized in the *Recommendations for Test Publishers* below. In addition, ETS researchers and other researchers concerned about these issues should focus on answering many of the questions raised by this study; these are summarized in the *Recommendations for Further Research* below.

Recommendations for Test Publishers

Because ETS procedures proved unable to identify sex-biased items on the two SATs studied (June 1986 and November 1987), different procedures are needed to reduce test bias:

1) ETS should eliminate from future SAT Verbal and Math tests those questions that have been identified by this analysis as showing the largest gender, race, and class differences (see Chapter 2). Removing items from the test that have large response differences between the sexes, unless they are balanced by other items, is a first step towards achieving balance and fairness without compromising test integrity. ETS and other test publishers should evaluate items after their inclusion on experimental sections, comparing the percentage correct between the sexes and among races and social classes to identify extreme items. However, deletion of items favoring one sex or ethnic group may not in itself make the SAT fairer, for a test might contain no such extreme items, yet still be biased against women or men or racial or ethnic groups. To put this another way, including male-biased items such as "mercenary" might be defensible, provided they are balanced with enough female-biased items such as "sheen." A fair test assembly procedure must first be aware of biased items and then use this knowledge to construct fair tests.

2) Since male and female mean scores on the verbal test are arbitrary and manipulable by the test-maker, the test-maker can manipulate them so that males and females score equally well, based on ability and knowledge; this would contribute to development of a sex-equal verbal test. Areas where women and girls have been shown to excel, such as writing and human relations, either are not evaluated by the SAT or are downplayed in favor of math, science, and business items. The SAT should test a more balanced array of skills and knowledge.

3) ETS and other test publishers should publicize the validity studies they

now conduct on the relationship between SAT scores and first year college grades and should make their findings available not only to other researchers but also to consumers—so that parents, students, and colleges, for example, can determine how much weight to give SAT scores as compared to high school grades and other evaluation factors as predictors of first year college performance.

4) ETS and other test publishers also should perform more research correlating performance on each SAT question with college grades. Such research should lead to future SATs with higher correlations to first year college grades, which would probably reduce test bias and certainly increase test defensibility.

5) ETS and other test publishers should allow test takers more time for each section of the test, to overcome the problems inherent in speeded tests, especially for women and girls. Although additional research is needed to further assess the impact of time pressure on girls and on students of color, enough evidence exists to indicate that the performance of women of all racial/ethnic groups and men of color would improve if they were allowed more time to complete the test.

Recommendations for Further Research

1) Conduct research on the predictive validity of the SAT and ACT for the college performance of women and men of color, including African American, Asian American, Hispanic, and Native American students of all socioeconomic levels. Such research could assess the nature and extent of race-plus-sex bias in standardized tests and determine whether it is possible to create a sex and race/culture-fair test.

2) Investigate the connections between sex and race bias in the classroom (including teacher-student interaction, textbooks, and other educational experiences) and bias in testing, to further assess the extent to which the SAT measures and therefore values the skills and knowledge that still differentiate upper middle class white males from others.

3) Conduct research on the impact of coaching on women and girls, students of color, and low income students. The success of Princeton Review and other coaching schools would seem to indicate that coaching is successful in enabling its students to improve their SAT scores. Further research might contribute to knowledge of how "playing the game" and "learning the tricks" of the SAT contribute to SAT success and thus could help the test makers to develop improved standardized tests.

4) Conduct further research on test anxiety to investigate further why girls and upwardly mobile boys are more anxious, so that steps can be taken to decrease their anxiety. However, if students accurately perceive that their educational futures are at stake and also believe that they test below their true ability, test anxiety may not be ameliorable by programs based on research. In addition, research that identifies this or other causes of test anxiety and assesses its impact on test performance could lead to recommendations that would assist test publishers, colleges, parents and teachers to consider alternatives to over-reliance on standardized testing.

At minimum, since girls exhibit more anxiety about the Math SAT than boys, score worse on it, and yet do as well or better in math courses, further research could determine whether girls' anxiety could be reduced if the verbal content of more math items were changed to favor girls.

5) Conduct further research to explain one of this study's most surprising and distressing findings: that the largest sex differences in SAT score

averages—much larger than the national averages for the test as a whole—occurred between boys and girls with the highest high school grades (A+ to A), while the smallest gender gap occurred at the lowest GPA level. Additional research with this population could find alternative explanations to the standard explanation offered by ETS—that the score gap occurs as a result of differential course-taking by male and female students—an explanation that already has been challenged by research in Montgomery County, Maryland (Gross, 1988).

6) Conduct research and development efforts that would contribute to the development of useful, predictive, and fair alternatives to standardized testing to evaluate students' achievements and predict their future performance. Following the lead of MIT, Bowdoin, Bates and others, for example, colleges could begin this process by conducting their own predictive validity studies and assessing their own need for the SAT to select their students.

For example, research could assess the predictive validity of different combinations of assessment instruments, including SAT and ACT scores, high school grades, class rank, and other tools that are used in evaluation. These may include writing samples done in a testing environment, various test scores such as achievement tests that cover specific curricula in a number of subject areas, and teachers' responses to questionnaires that ask a variety of questions about their students' skills, knowledge, talents and potential.

Complete Bibliography and References

Adams, Raymond J. 1986. "Some Contributions to Sex Differences in Scholastic Aptitude Scores." *Studies in Educational Evaluation* 12(3):267-74.

Adams, Raymond J. 1985. "Sex and Background Factors: Effect on ASAT Scores." *Australian Journal of Education*. 29(3):221-30, November.

Alderman, Donald. 1981. "Student Self-selection and Test Repetition." *Educational and Psychological Measurement* 41(4):1073-81.

Allison, Donald E. 1984. "The Effect of Item-difficulty Sequence, Intelligence, and Sex on Test Performance, Reliability, and Item Difficulty and Discrimination." *Measurement and Evaluation in Guidance* 16(4):211-217.

American Civil Liberties Union. 1988. *Affidavits for Khadijah Sharif, et. al. v. New York State Education Department, et. al. Civ. 8435.*

American College Testing Program (ACT). 1973. *Assessing Students on the Way to College*. Iowa City: American College Testing Program.

ACT. 1987. *College Student Profiles: Norms for the ACT Assessment*. Iowa City: American College Testing Program.

American Council on Education. 1985. "Minorities in Higher Education." *Fourth Annual Status Report*. Washington, D.C.

Anastasi, A. 1981. "Sex Differences: Historical Perspectives and Methodological Implications." *Developmental Review*, 1:187-206.

Angoff, William H. 1988. "The Nature-Nurture Debate, Aptitudes, and Group Differences." *American Psychologist*. Vol. 43, pp. 95-103, February.

Anrig, Gregory. 1987. "'Golden Rule:' Second Thoughts." *American Psychological Association Monitor*, 3 January.

Ash, Philip. 1966. "The Implications of the Civil Rights Act of 1964 for Psychological Assessment in Industry." *American Psychologist* 21(8):797-803.

Associated Press (AP). 1987. Drug May Help the Overanxious on S.A.T.s. *New York Times*, October 22, p. A27.

Becker, Betsy J. 1983. Item characteristics and sex differences on the SAT-M for mathematically-able youths. Paper presented at the annual meeting of the American Educational Research Association, May; Montreal, Canada.

Behnke, Michael, J. 1987. Washington, D.C.: Testimony Presented to the Congressional Subcommittee on Civil and Constitutional Rights. April 23.

Bem, S.L. and Bem, D.V. 1970. "We're All Nonconscious Sexists." *Psychology Today*, April.

Benbow, Camilla and Julian C. Stanley. 1983a. "Differential Course-taking Hypothesis Revisited." *American Educational Research Journal* 20(4):469-73.

Benbow, Camilla and Julian C. Stanley. 1983b. "Sex Differences in Mathematical Reasoning Ability: More Facts." *Science* 222(4627):1029-30.

Benbow, Camilla and Julian C. Stanley. 1982. "Consequences in High School and College of Sex Differences in Mathematical Reasoning Ability: A Longitudinal Perspective." *American Educational Research Journal* 19(4):598-622.

Benbow, Camilla and Julian C. Stanley. 1980. "Sex Differences in Mathematical Ability: Fact or Artifact?" *Science* 210(4475):1262-64.

Billingham, Katherine A., *et. al.* 1981. Influences of sex differences and achievement on test anxiety. Paper presented at the annual meeting of the American Psychological Association, 24-26 August, Los Angeles.

Bolger, Niall. 1984. Gender differences in academic achievement according to method of measurement. Paper presented at the annual meeting of the American Psychological Association, August, Toronto, Ontario.

Bouldt, Robert F. 1983. *Status of Research on Item Content and Differential Performance on Tests Used in Higher Education*. Princeton: Educational Testing Service.

Boyer, Ernest L. 1987. *College: The Undergraduate Experience in America*. New York: Harper and Row.

Brandt, Ron. 1980. "On Admissions Testing: A Conversation With Fred Hargadon." *Educational Leadership* 37(8):655-657.

Breland, Hunter M. and Philip A. Griswold. 1982a. "Use of a Performance Test as a Criterion in a Differential Validity Study." *Journal of Educational Psychology* 74(5):713-21.

Breland, Hunter M. and Philip A. Griswold. 1982b. *Group Comparison for Basic Skills Measures*. (College Board Report No. 81-6). Princeton: Educational Testing Service.

Breland, Hunter M. 1978. *Population Validity and College Entrance Measures*. (Research and Development Report 78-79, No. 2) New York: College Entrance Examination Board.

Brown, Fred G. 1980. "Sex Bias in Achievement Test Items: Do They Have Any Effect on Performance?" *Teaching of Psychology* 7(1):24-6.

Burton, Nancy W. 1987. Trends in the verbal scores of women taking the SAT in comparison to trends in other voluntary testing programs. Paper presented at the annual meeting of the American Educational Research Association, April, Washington, D.C.

Burton, Nancy W. and Charles Lewis. 1988. Modelling women's performance on the SAT. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Calkins, Rick S., and Randolph Whitworth. 1974. Differential prediction of freshman grade point average for sex and two ethnic classifications at a southwestern university. ERIC Report.

Campbell, Patricia, B. 1981. *The Impact of Societal Biases on Research Methods*. Washington D.C.: National Institute of Education.

Cannell, John J. 1987. "Nationally-Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average." Daniels, West Virginia: Friends for Education.

Chipman, Susan. 1988. Word problems: Where test bias creeps in. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Clark, Mary Jo, and Jerilee Grandy. 1984. *Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers.* Report No. 84-8. New York: College Entrance Examination Board.

Coffman, William E. 1961. "Sex Differences in Response to Items in an Aptitude Test." *National Council on Measurement in Education, Eighteenth Yearbook.* Michigan. pp. 117-124.

Cole, Nancy C. 1981. "Bias in Testing." *American Psychologist* 36:1067-77.

Cole, Nancy C. 1973. "Bias in Selection." *Journal of Educational Measurement* 10(2):237-255.

The College Board. 1988. *1987-88 ATP Guide for High Schools and Colleges.* New York: College Entrance Examination Board.

The College Board. 1988. *5 SATs 1988 Edition.* New York: CEEB.

The College Board. 1988. *News From the College Board.* New York: The College Board, September 20.

The College Board. 1988. *National College-Bound Seniors: 1988 Profiles of SAT and Achievement Test Takers.* New York: CEEB.

The College Board. 1987. *1986-87 ATP Guide for High Schools and Colleges,* New York: CEEB.

College Entrance Examination Board. 1988. *1988 National Ethnic/Sex Profiles.* (for the SAT) New York: CEEB.

Commission on Sex Bias in Measurement. 1977. "A Case History of Change: A Review of Responses to the Challenge of Sex Bias in Career Interest Inventories." *Measurement and Evaluation in Guidance* 10(3):143-152.

Cordes, Colleen. 1986. "Test Tilt: Boys Outscore Girls on Both Parts of SAT." *Monitor,* American Psychological Association, June.

Costa, Arthur, ed. 1985. *Developing Minds.* Alexandria, VA: Association of Supervision and Curriculum Development.

Covert, Robert W., and Norman M. Chansky. 1975. "The Moderator Effect of Undergraduate Grade Point Average on the Prediction of Success in Graduate Education." *Educational and Psychological Measurement* 35(4):947-50.

Crouse, James and Dale Trusheim. 1988. *The Case Against the SAT.* Chicago: The University of Chicago Press.

Dappen, Leon, and Cecil R. Reynolds. 1981. "Factorial Validity of the 1976 Edition of the Metropolitan Readiness Tests for Males and Females." *Psychology in the Schools* 18(4):413-416.

Datta, Lois-Ellin. 1977. *He and She: Sex Fairness in Selection and Guidance Based on Testing.* Washington D.C.: National Institute of Education.

deNys, Mary, and Leslie R. Wolfe. 1985. *Learning Her Place—Sex Bias in the Elementary School Classroom*. Washington, DC: Project on Equal Education Rights.

Diamond, Esther E. 1985. Content, context and construct considerations in sex bias in testing. Paper presented at the annual meeting of the American Educational Research Association, 31 March-4 April, Chicago.

Diamond, Esther E. 1976. "Minimizing Sex Bias in Testing." *Measurement and Evaluation in Guidance* 9(1):28-34.

Diamond, Esther E., and Patricia B. Elmore. 1986. "Bias in Achievement Testing: Follow-up Report of the AMECD Commission on Bias in Measurement." *Measurement and Evaluation in Counseling and Development* 19(2):102-112. July.

Diamond, Esther E. and Carol Kehr Tittle. 1985. "Sex Equity in Testing." in Klein, Susan, ed. *Handbook for Achieving Sex Equity Through Education*. Baltimore: Johns Hopkins University Press.

Dong, Hei-ki, Young H. Sung and Thomas E. Dohm. 1985. "The Validity of the Ball Aptitude Battery (BAB): In Relationship to High School Academic Success." *Educational & Psychological Measurement* 45(3):627-637.

Donlon, Thomas F. ed. 1984. *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: CEEB.

Donlon, Thomas F. 1977. Sex differences in test speededness on the SAT. Paper presented at the annual meeting of the Northeastern Educational Research Association, 26-28 October, Ellenville, New York.

Donlon, Thomas F., et al. 1977. *Performance Consequences of Sex Bias in the Content of Major Achievement Batteries. Final Report*. Princeton: Educational Testing Service.

Donlon, Thomas F. 1973. *Content Factors in Sex Differences on Test Questions*. Research Memorandum 73-28, Princeton: Educational Testing Service.

Donlon, Thomas F., and William H. Angoff. 1971. "The Scholastic Aptitude Test." *The College Board Admissions Testing Program*, William H. Angoff, Ed., pp. 15-47, New York: College Entrance Examination Board.

Donlon, Thomas F., Ruth B. Ekstrom and Marlaine E. Lockheed. 1979. "The Consequences of Sex Bias on the Content of Major Achievement Test Batteries." *Measurement and Evaluation in Guidance*, 11(4):202-216. January.

Donlon, Thomas F., Marilyn M. Hicks, and Madeline M. Wallmark. 1980. "Sex Differences in Item Responses on the Graduate Record Examination." *Applied Psychology Measurements* 4(1):9-20.

Doolittle, Allen E. 1985. Understanding differential item performance as a consequence of gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Assoication, 31 March-4 April, Chicago.

Doolittle, Allen E. 1987. Gender Differences in Performance on Mathematics Achievement Items. Paper presented at the annual meeting of the American Psychological Association, August, New York.

Dorans, Neil J., and Edward Kulick. 1983. *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach*. Princeton: Educational Testing Service.

Dorans, Neil J., Alicia P. Schmitt and W. Edward Curley. 1988. *Differential Speededness: Some Items Have DIF Because of Where They Are, Not What They Are*. Princeton: Educational Testing Service.

Durio, Helen F., *et. al.* 1980. Ethnicity and sex differences in use of college examination, mathematics achievement, and high school rank as predictors of performance and retention among engineering students. Paper presented at the annual meeting of the American Educational Research Association, 7-11 April, Boston.

Dwyer, Carol A. 1979. "The Role of Tests and Their Construction in Producing Apparent Sex-related Differences." *Sex Related Differences in Cognitive Functioning*, Wittig and Petersen, eds. New York: Academic Press.

Dwyer, Carol A. 1976a. "Test Content and Sex Differences in Reading." *The Reading Teacher* 29(8):753-77.

Dwyer, Carol A. 1976b. Test content in mathematics and science: the consideration of sex. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

ETS Sensitivity Review Process. 1987. Princeton: Educational Testing Service.

Educational Research Service. 1981. *Testing for College Admissions: Trends and Issues*. Arlington: Educational Research Service.

Ekstrom, Ruth B., Marlaine E. Lockheed and Thomas F. Donlon. 1979. "Sex Differences and Sex Bias in Test Content." *Educational Horizons* 58(1):47-52.

Erkut, Sumru. 1983. "Exploring Sex Differences in Expectancy, Attribution, and Academic Achievement." *Sex Roles* 9(2):217-31.

Faggen-Streckler, Jane. 1974. "A Quantitative Method for Measuring Sex 'Bias' in Standardized Tests." *Journal of Educational Measurement* 11(3):151-61.

Fallows, James. 1980. "The Tests and the Brightest: How Fair Are The College Boards?" *Journal of the National Association of College Admissions Counselors* 24(3):14-31.

Feingold, Alan. 1988. "Cognitive Gender Differences Are Disappearing." *American Psychologist*, February, pp. 95-103.

Fendrich-Saloweg, Gail, Mary Bucharan, and Clifford J. Drew. 1982. "Mathematics, Quantitative and Attitudinal Measures for Elementary School Boys and Girls." *Psychological Reports* 51(1):155-62.

Fennema, Elizabeth L. and Julia Sherman. 1977. "Sex-related Differences in Mathematics Achievement, Spatial Visualization and Affective Factors." *American Educational Research Journal*, 14, pp. 51-71.

Ferber, Marianne A., *et al.* 1983. "Gender Differences in Economic Knowledge: A Re-evaluation of the Evidence. *Journal of Economic Education* 14(2):24-37.

Fitzgerald, Laurine E., and B. Jeanne Fisher. 1974. *Legal Issues: Status Report.* Minneapolis: Aries Corporation.

Flaugher, Ronald L. 1974. *Bias in Testing: A Review and Discussion.* TM Report No. 36. Princeton: Clearinghouse on Tests, Measurement, Evaluation.

Frank, Austin C., and Katharine M. Jeffrey. 1978. *Freshman Admission by Formula: A Retrospective Study of Impact on Student Mix and Graduation Rates at Berkeley.* Berkeley: UC Berkeley Office of Student Affairs Research.

Freed, Norman H. 1983. "Prospective Mathematical Equivalence by Gender: Still More Inadvertent Support." *Psychological Report* 53(2):677-678.

Fyans, Leslie J. Jr. 1979. Test anxiety, test comfort and student achievement test performance. Paper presented at ETS Seminar, Princeton.

Gamache, LeAnn M., and Melvin Novick. 1985. "Choice of Variables and Gender Differentiated Prediction Within Selected Academic Programs." *Journal of Educational Measurements* 22(1):53-70.

Gamache, LeAnn M., and Melvin Novick. 1983. *Choice of Variable and Gender Differentiated Prediction Within Selected Academic Programs.* Research Report No. 105. Iowa City: Iowa University Evaluation and Examination Service.

Gold, Ann. 1977. "The Use of Separate-sex Norms on Aptitude Tests: Friend or Foe?" *Measurement and Evaluation in Guidance* 10(3):162-70.

Graf, Richard G. and Jeanne C. Riddell. 1972. "Sex Differences in Problem Solving as a Function of a Problem Context." *Journal of Educational Research* 65(10):451-2.

Grandy, Jerilee. 1987. *Ten-Year Trends in SAT Scores and Other Characteristics of High School Seniors Taking the SAT and Planning to Study Mathematics, Science or Engineering.* Princeton: Educational Testing Service.

Grant, Marvin L., and Roy Singleton, Jr. 1982. "Alternative Admissions Criteria in the 80's." *Negro Educational Review* 33(3-4):167-75.

Green, Donald Ross. 1987. Sex differences in item performance on a standardized achievement battery. Paper presented at the annual meeting of the American Psychological Association, New York.

Gross, Susan. 1988. *Participation and Performance of Women and Minorities in Mathematics.* Maryland: Department of Educational Accountability, Montgomery County Schools. July.

Hackett, Rachelle Kisst, Paul Holland, Mari Pearlman and Dorothy Thayer. 1987. "Test Construction Manipulating Score Differences Between Black and White Examinees: Properties of the Resulting Tests." Princeton: Educational Testing Service. February.

Hale, Robert L. and Audrey A. Potok. 1980. "Sexual Bias in the WISC-R." *Journal of Consulting and Clinical Psychology* 48(6):776.

Hamilton, Lawrence C. 1981. "Sex Differences in Self-report Errors: A Note of Caution." *Journal of Education Measurement* 18(4):221-8.

Harris, Abigail M. 1976. *ETS Studies Related to Women and Education: Annotated Bibliography.* Princeton: Educational Testing Service.

Herman, David O. 1981. Reducing sex bias in ability tests. Paper presented at the annual meeting of the American Psychological Association, August, Los Angeles.

Hoffman, L. Richard, and Norman R.F. Maier. 1966. "Social Factors Influencing Problem Solving in Women." *Journal of Personality and Social Psychology* 4(4):282-90.

Hoge, Robert D., and Robert Butcher. 1984. "Analysis of Teacher Judgments of Pupil Achievement Levels." *Journal of Educational Psychology* 76(5):777-81.

Hogrebe, Mark C., *et al.* 1983. "The Moderating Effects of Gender and Race in Predicting the Academic Performance of College Development Students." *Educational and Psychological Measurement* 43(2):523-30.

Holland, J.L. and R.L. Nichols. 1964. "Prediction of Academic and Extra-curricular Achievement in College. *Journal of Educational Psychology* 55:55-65.

Hoover, Mary Rhodes, Robert L. Politzer and Orlando Taylor. 1987. "Bias in Reading Tests for Black Language Speakers: A Sociolinguistic Perspective." *The Negro Educational Review Special Issue on Testing African American Students* Vol. 38, Nos. 2-3, pp. 67-80, April-July.

Horner, Blair and Joe Sammons. 1988. *Rolling Loaded Dice.* New York: New York Public Interest Research Group, Inc.

Hunt, Barbara. 1979. *Sex Bias in Testing: An Annotated Bibliography.* Washington D.C.: National Institute of Education.

Hyde, Janet Shibley and Marcia C. Linn. 1988. "Gender Differences in Verbal Ability: A Meta-Analysis." *Psychological Bulletin* 104(1):53-69

Jacobs, J.E., and Jacquelynne S. Eccles. 1985. "Gender Differences in Math Ability: The Impact of Media Reports on Parents." *Educational Researcher* 14,3:20-24.

Jacobson, Robert L. 1986. "Selective Colleges' Use of SAT is Unshaken by Controversies." *Chronicle of Higher Education* 32(18).

Johnson, Sylvia T. 1979. *The Measurement Mystique: Issues in Selection for Professional Schools and Employment.* Occasional Paper No. 2-79. Washington D.C.: Howard University, Institute for the Study of Educational Policy.

Jones, Robert F. and Suzanne Vanyur. 1985. An investigation of gender-related test bias for the medical college admission test. Paper presented at the annual meeting of the National Council of Measurement in Education, 1-4 April, Chicago.

Kanarek, Ellen. 1988. Gender differences in freshman performance and their relationship to use of the SAT in admissions. Paper presented at the annual meeting of the Northeast Association for Institutional Research, October, Providence, RI.

Karmos, Ann H. and Joseph S. Karmos. 1984. "Attitudes Toward Standardized Achievement Tests and Their Relation to Achievement Test Performance." *Measurement and Evaluation in Counseling and Development* 17(2):56-66.

Kirschenbaum, Robert J. 1983. "Let's Cut Out the Cut-Off Score in the Identification of the Gifted." *Roeper Review* 5(4):6-10.

Klein, Susan S. 1986. Why do females receive higher course grades, but often lower standardized achievement and aptitude test scores than males? Paper presented at the National Center for Fair and Open Testing Conference, 11 December, Washington, D.C.

Klein, Susan S., ed. 1985. *Handbook for Achieving Sex Equity in Education.* Baltimore: Johns Hopkins University Press.

Kleinke, David J. 1980. "Item, Order, Response Location and Examinee Sex and Handedness and Performance on a Multiple Choice Test." *Journal of Educational Research* 73(4):225-229.

Kulick, Edward and Neil J. Dorans. 1984. The standardization approach to assessing unexpected differential item performance. Paper presented at the annual meeting of the American Educational Research Association, 23-27 April, New Orleans.

Larson, James R., and Peter M. Scontrino. 1976. "The Consistency of High School Grade Point Average and of the Verbal and Mathematical Portions of the SAT of the CEE Board, as Predictors of College Performance: An Eight Year Study." *Educational and Psychological Measurement* 36(2):439-43.

Lawrence, Ida M., W. Edward Curley and Frederick J. McHale. 1988. Differential item functioning of SAT-Verbal reading subscore items for male and female examinees. Paper presented at the annual meeting of the National Council on Measurement in Education and the American Educational Research Association, April, New Orleans.

Linn, Marcia C. and Janet S. Hyde. 1988. Gender, mathematics, and science. Unpublished paper, June.

Linn, Marcia C., Tina DeBenedictis, Kevin Delucchi, Abigail Harris and Elizabeth Stage. 1987. "Gender Differences in National Assessment of Educational Progress Science Items: What Does 'I Don't know' Really Mean?" *Journal of Research in Science Teaching* Vol. 24, No. 3, pp. 267-278.

Linn, Robert L. 1982. Selection bias: multiple meanings. Presidential Address to the Division of Evaluation and Measurement at the annual meeting of the American Psychological Association, August.

Linn, Robert L. 1978. "Single-group Validity, Differential Validity, and Differential Prediction." *Journal of Applied Psychology* 63(4):507-12.

Linn, Robert L. 1973. "Fair Test Use in Selection." *Review of Educational Research* 43:139-61.

Lockheed, Marlaine E. 1982. "Sex Bias in Aptitude and Achievement Tests Used in Higher Education." *The Undergraduate Woman: Issues in Educational Equity,* P.J. Perun, Ed. Lexington: Lexington Books.

Lockheed, Marlaine E. 1974a. *Sex Discrimination in Education: A Literature Review and Bibliography.* Princeton: Educational Testing Service.

Lockheed, Marlaine E. 1974b. "Sex Bias in Educational Testing: A Sociologist's Perspective." *Research Memorandum* No. 74-13. Princeton: Educational Testing Service.

Loewen, James W. 1979. "Introductory Sociology: Four Classroom Exercises." *Teaching Sociology,* Volume 6, No. 3 (April), pp. 221-244.

Loewen, James W., Phyllis Rosser, and John Katzman. 1988. Gender bias in SAT items. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Logan, Samuel H. 1980. "Testing for Sex Bias in Graduate School Admissions." *College and University* 55(2):156-70.

Lueptow, Lloyd B. 1980. "Gender Wording, Sex, and Response to Items on Achievement Value." *Psychological Reports* 46(1):140-2.

Maccoby, Eleanor E. and Carol N. Jacklin. 1974. *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.

McCarthy, K. 1975. Sex bias in tests of mathematical aptitude. Doctoral dissertation, City University of New York. New York.

McCarty, Joyce R., Candace Noble and Renee Huntley. 1988. Effects of item wording on sex bias. Paper presented at the annual meeting of the National Council on Measurement in Education, April, New Orleans.

McCornack, Robert L. 1983. "Bias in the Validity of Predicted College Grades in Four Ethnic Minority Groups." *Educational and Psychological Measurement* 43(2):517-22.

McPeek, W. Miles and Cheryl Wild. 1987. Characteristics of quantitative items that function differently for men and women. Paper presented at the annual meeting of the American Psychological Association, August, New York.

Medley, Donald M. and Thomas J. Quirk. 1974. "The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teachers Examination (NTE)." *Journal of Educational Measurement* 2(4) Winter.

Meyer, Paul R. 1982. A study of sex differences in the freshman composition course at the University of Texas at Austin. ERIC Reports.

Miller, Arden, and William Asher. 1983. "Gender Differences as a Factor in Teachers' Perceptions of Students: A Comment." *Perception and Motor Skills* 56(3):856-58.

Milton, G. A. 1959. "Sex differences in problem solving as a function of role appropriateness of problem content." *Psychological Reports* 5(4):705-8.

Milton, G. A. 1958. "Five Studies of the Relation Between Sex Role Identification and Achievement in Problem Solving," Technical Report No. 3, Department of Industrial Administration, Department of Psychology. New Haven, CT: Yale University, December.

Milton, G. A. 1957. "The Effects of Sex Role Identification Upon Problem Solving Skill." *Journal of Abnormal and Social Psychology*, Volume 55, pp. 208-212.

Minatoya, Lydia, and William E. Sedlacek. 1981. *The SASW: A Measure of Sexism Among University Freshman*. College Park: Maryland University Counseling Center.

Mullis, Ina V. S. 1987. Trends in performance for women taking the NAEP reading and writing assessments. Paper presented at the annual meeting of the American Educational Research Association, April, Washington, DC.

Murphy, R.J. 1982. "Sex differences in Objective Test Performance." *British Journal of Educational Psychology* 52(2):213-19.

Novick, M.R. 1982. "Educational Testing: Inferences in Relevant Subpopulations." *Educational Researcher* II:4-10.

Pallas, Aaron M., and Karl L. Alexander. 1983. "Sex Differences in Quantitative SAT Performance: New Evidence on the Differential Coursework Hypothesis." *American Educational Research Journal* 20(2):165-82.

Parsons, Jacquelynne E., Terry Adler, and Judith L. Meece. 1984. "Sex Differences in Achievement: A Test of Alternate Theories." *Journal of Personality and Social Psychology* 46(1):26-43.

Paulhaus, Delroy, and David R. Schafter. 1981. "Sex Differences in the Impact of Number of Older and Number of Younger Siblings on Scholastic Aptitude." *Social Psychology Quarterly* 44(4):363-8.

Payne, Beverly D. 1984. "The Relationship of Test Anxiety and Answer-changing Behavior: An Analysis by Race and Sex. *Measurement and Evaluation in Guidance* 16(4):205-10.

Payne, Beverly D., Janet E. Smith, and David A. Payne. 1983. "Grades, Sex and Race Differences in Test Anxiety." *Psychological Reports* 53(1):291-4.

Pfeifer, C. Michael Jr., and William E. Sedlacek. 1971. "The Validity of Academic Predictors for Black and White Students at a Predominantly White University." *Journal of Educational Measurement* 8(4):253-61.

Plake, Barbara S., Charles J. Ansorge, Claire S. Parker and Steven R. Lowry. 1982. "Effects of Item Arrangement, Knowledge of Arrangement, Test Anxiety and Sex on Test Performance." *Journal of Educational Measurement* 19(1):49-57. Spring.

Plake, Barbara S., H. D. Hoover, and Brenda H. Loyd. 1981. "Sex Differences in Mathematics Components of the Iowa Tests of Basic Skills." *Psychology of Women Quarterly* 5 (5, suppl.):780-4.

Plake, Barbara S., H. D. Hoover, and Brenda H. Loyd. 1980. "An Investigation of the Iowa Tests of Basic Skills for Sex Bias: A Developmental Look." *Psychology in the Schools* 17(1):47-52.

Plake, Barbara S., H. D. Hoover, and Brenda H. Loyd. 1978. An investigation of differential item performances by sex on the Iowa Tests of Basic Skills. Paper presented at the annual meeting of the National Council of Measurement and Education, March, Toronto.

Powell, Brian, and Lala Carr Steelman. 1982. "Testing for Sex Inequality in Standardized Admissions Exams: The Case for Open Access." *Integrated Education* 20(3-5):86-8.

Powers, Stephen, Cathy Escamilla and Myra M. Haussler. 1986. "The California Achievement Test as a Predictor of Reading Ability across Race and Sex." *Educational and Psychological Measurement* 46(4):1067-70.

Powers, Stephen, and Patricia D. Jones, *et. al.* 1984. "Factorial Invariance of the California Achievement Tests Across Race and Sex." *Educational and Psychological Measurement* 44(4):967-75.

Powers, Stephen, Douglas Thompson, Barbara Azevedo, and Olivia Schaad *et. al.* 1983. "The Predictive Validity of the Stanford Mathematics Test Across Race and Sex. *Educational and Psychological Measurement* 43(2):645-9.

Ramist, Leonard, and Solomon Arbeiter. 1986. *Profiles, College-Bound Seniors, 1985*. New York: College Entrance Examination Board.

Rock, Donald A., Randy Elliot Bennett, and Bruce A. Kaplan. 1985. "The Internal Construct Validity of the SAT Across Handicapped and Non-handicapped Populations. *College Board Report* No.4.

Rogers, Brenda. 1984. The use of non-cognitive variables in the prediction of black freshmen's academic performance. Paper presented at the annual meeting of the Southern Association for Institutional Research.

Roid, Gale H., and Cathy L.W. Wendler. 1983. Item bias detection and item writing technology. Paper presented at the annual meeting of the American Educational Research Association, 11-15 April, Montreal.

Rosser, Phyllis. 1987. *Sex Bias in College Admissions Tests: Why Women Lose Out.* Cambridge, Massachusetts: The National Center for Fair and Open Testing.

Rowell, E.H. 1977. *Are Reading Tests Sexist? An Investigation into Sex Bias in Three Frequently Used Individualized Reading Tests.* Providence: Rhode Island College.

Sadker, Myra, and David Sadker. 1986. "From Grade School to Graduate School: Sex Bias in Classroom Interaction." *Phi Delta Kappan,* April.

Sadker, Myra, and David Sadker. 1985. "Sexism in the Schoolroom of the 80s." *Psychology Today,* March.

Scheuneman, Janice Dowd and Jacqueline B. Briel. 1988. Differential effects of selected item features on the performance of Hispanic and white examinees. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Scheuneman, Janice Dowd. 1987. "An Experimental Exploratory Study of Causes of Bias in Test Items." *Journal of Educational Measurement* 24(2):97-118.

Schmitt, Alicia P. 1988. "Language and Cultural Characteristics That Explain Differential Item Functioning for Hispanic Examinees on the Scholastic Aptitude Test." *Journal of Educational Measurement* Vol.25, No.1, pp. 1-13, Spring.

Schmitt, Alicia P., and Carole A. Bleistein. 1987. *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items.* Princeton: Educational Testing Service.

Schmitt, Alicia P., W. Edward Curley, Carole A. Bleistein and Neil J. Dorans. 1988. *Experimental Evaluation of Language and Interest Factors Related to Differential Item Functioning for Hispanic Examinees on the SAT-Verbal.* Princeton: Educational Testing Service.

Schmitt, Alicia P. and Neil J. Dorans. 1987. Differential item functioning for minority examinees on the SAT. Paper presented at the annual meeting of the American Psychological Association, August, New York.

Schofield, Hilary L. 1982. "Sex, Grade Level, and the Relationship Between Mathematics Attitude and Achievement in Children." *Journal of Educational Research* 75(5):280-4.

Schonberger, Ann K. 1978. Are mathematics problems a problem for women and girls? Paper presented at the annual meeting of the National Council of Teachers of Mathematics, April, San Diego.

Schrader, William B. 1984. *Three Studies of SAT-Verbal Types.* College Board Report No. 84-7. New York: CEEB.

Schratz, Mary K., and Barrie Wellens. 1981. Minority panel review in the development of an achievement test. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Sedlacek, William E. 1976. *Recent Developments in Test Bias Research.* University of Maryland Study Center Research Report No. 2-76. College Park: University of Maryland Cultural Study Center.

Selkow, Paula. 1984. *Assessing Sex Bias in Testing: A Review of the Issues and Evaluations of 74 Psychological and Educational Tests.* Westport: Greenwood Press.

Sinnot, Lorraine. 1982. *The Identification of Biased Items.* Princeton: Educational Testing Service.

Smith, Richard M. 1988. Differential item familiarity in graduate admissions mathematics tests. Paper presented at the annual meeting of the National Council on Measurement in Education, April, New Orleans.

Smith, Richard M. 1983. Test fairness is a personal issue. Paper presented at the annual meeting of the National Council on Measurement in Education, 12-14 April, Montreal.

Smith, Richard M. 1981. An analysis of individual effects of sex bias. Paper presented at the annual meeting of the National Council on Measurement in Education, 13-17 April, Los Angeles.

Speth, Carol A., and Barbara S. Plake. 1985. Assessment of positive sex-role characteristics. Paper presented at the annual meeting of the American Psychological Association, 23-27 August, Los Angeles.

Sranum, Soren, and Robert G. Bringle. 1982. "Race, Social Class, and Predictive Bias: An Evaluation Using the WISC, WRAT, and Teacher Ratings." *Intelligence* 6(3):275-86.

Stage, Elizabeth K. 1986. Affective and attitudinal factors associated with the performance of female and minority students in mathematics and science. Paper presented at the annual meeting of the American Association for the Advancement of Science, May, Philadelphia.

Stanley, Julian. 1986. *SAT-M Scores of Highly Selected Students in Shanghai Tested When Less than 13 Years Old.* (College Board Review No. 140).

Stanley, Julian, and Camilla Benbow. 1983. "15:1 Isn't "Catching Up!" *Psychological Reports* 52:656.

Stanley, Julian, and Camilla Benbow. 1983. "SMPY's First Decade: Ten Years of Posing Problems and Solving Them." *Journal of Special Education* 17(1):11-25.

Stoddard, Ann H. 1984. "Intelligence Testing Revisited." *Negro Educational Review* 35(1):17-24.

Strassberg-Rosenberg, Barbara, and Thomas F. Donlon. 1975. Content influences on sex differences in performance on aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, 31 March-2 April, Washington, D.C.

Stricker, Lawrence J. 1982. "Identifying Test Items that Perform Differentially in Population Subgroups: A Partial Correlation Index." *Applied Psychological Measurement* 6(3):261-73. Summer.

Taylor, Orlando and Dorian Latham Lee. 1987. "Standardized Tests and African Americans: Communication and Language Issues." *The Negro Educational Review Special Issue on Testing African American Students* Vol. 38, Nos. 2-3, pp. 67-80. April-July.

Thomas, Charles L. 1973. The overprediction phenomenon among Black collegians: some preliminary considerations. Paper presented at the annual meeting of the American Educational Research Association, 25 February-1 March, New Orleans.

Thomas, Gail E. 1986. "Cultivating the Interest of Women and Minorities in High School Mathematics and Science." *Science Education* 70(1):31-43.

Thomas, Hoben. 1985. "A Theory of High Mathematical Aptitude." *Journal of Mathematical Psychology* 29(2):231-42.

Tittle, Carol Kehr. 1982. "Use of Judgmental Methods in Item Test Bias Studies." R. Berk, ed. *Handbook of Methods for Detecting Test Bias.* Baltimore: Johns Hopkins University Press.

Tittle, Carol Kehr. 1978. *Sex Bias in Testing: A Review with Policy Recommendations.* San Francisco: Women's Educational Equity Communications Network, Far West Laboratory.

Tittle, Carol Kehr. 1974. *Women and Educational Testing: A Selective Review of the Research Literature and Testing Practices.* Princeton: Educational Testing Service.

U.S. District Court, Southern District of New York. *Preliminary Injunction Hearing Transcript,* January 23, 1989.

Van Duesen, William and Hal Higginbotham. 1984. *The Financial Aid Profession at Work: A Report on the 1983 Survey of Undergraduate Need Analysis Policies, Practices, and Procedures.* New York: CEEB.

Waetjen, Walter B. 1977. Sex differences in learning: some open questions. Paper presented at the annual meeting of the International Society for Study of Behavioral Development, 19-25 September, Pavia, Italy.

Warner, Howard. 1981. "Five Pitfalls Encountered While Trying to Compare States on their SAT Scores." *Journal of Educational Measurement* 23(1):69-81.

Wedman, Ingemar, and Christina Stage. 1983. "The Significance of Contents for Sex Differences in Test Results." *Scandinavian Journal of Educational Research* 27(1):49-71.

Welsh, Catherine J., and Allen E. Doolittle. 1988. Gender-based differential item performance in English Usage items. Paper presented at the annual meeting of the American Educational Research Association, 5-9 April, New Orleans.

Welsh, W.W., R.E. Anderson and L.J. Harris. 1982. "The Effects of Schooling on Mathematics Achievement." *American Educational Research Journal* 19:145–53.

Wendler, Cathy L.W. and Sydell T. Carlton. 1987. An examination of SAT Verbal items for differential performance by women and men: an exploratory study. Paper presented at the annual meeting of the American Educational Research Association, April, Washington, D.C.

Wild, Cheryl L., Robin Durso, and Donald B. Rubin. 1982. "Effect of Increased Test-taking Time on Test Scores by Ethnic Group, Years Out of School, and Sex." *Journal of Educational Measurement* 19(1):19–28.

Wild, Cheryl L., and Carol A. Dwyer. 1980. "Sex Bias in Selection." *Psychometrics for Educational Debate,* L.J. van der Kamp, N.M. de Gruijter and W.F. Langerak, Eds. New York: Wiley.

Wildemuth, Barbara M., comp. 1977. *Test Anxiety: An Extensive Bibliography.* Princeton: Clearinghouse on Tests, Measurement, and Evaluation.

Willingham, Warren W. "Handicapped Applicants to College: An Analysis of Admissions Decisions. *College Board Report* No. 87:1.

Wing, H. 1981. Practice effects with traditional mental test items: a replication. Paper presented at the annual meeting of the National Council for Measurement in Education.

Wirtenberg, T. Jeana, and Charles Y. Nakamura. 1976. "Educational Barrier or Boon to Changing Occupational Roles of Women?" *Journal of Social Issues* 32(3):165–79.

Zajonc, Robert B. 1986. "The Decline and Rise of Scholastic Aptitude Scores: A Prediction Derived from the Confluence Model." *American Psychologist* 41(8):862–7.

Zoref, Leslie, and Paul Williams. 1980. "A Look at Content Bias." *Journal of Educational Measurement* 17(4):313–22.

Zorn, Jeffrey L. 1983. Possible sources of culture bias in the validation of ETS Language Tests. Paper presented at the annual meeting of the Conference on College Composition and Communication, 17–19 March, Detroit.

Appendices

Appendix A: Items with Extreme Differences by Sex (June 1986 SAT)

Section 1

1. SETBACK:
   (A) commotion
   (B) variation
   (C) eagerness
   (D) concentration
   (E) employment

5. SHEEN:
   (A) uneven in length
   (B) dull finish
   (C) strong flavor
   (D) narrow margin
   (E) simple shape

23. The author's tone can best be described as which of the following?
   (A) Whimsical
   (B) Confidential
   (C) Narrative
   (D) Instructive
   (E) Speculative

44. MERCENARY:SOLDIER::
   (A) censor:author
   (B) hack:writer
   (C) agent:performer
   (D) fraud:artist
   (E) critic:subject

Section 4

21. PENDANT:JEWELRY::
   (A) frame:picture
   (A) cue:drama
   (C) violin:music
   (D) mobile:sculpture
   (E) poetry:prose

24. LOVE:REQUITE::
(A) attack:retaliate
(B) proposal:write
(C) problem:worry
(D) film:review
(E) law:domineer

31. Perrot betrays Wilson by revealing that
(A) Dawson's presence should be no surprise to Wilson
(B) Perrot's wife had expected Wilson's arrival
(C) Wilson has ignored the plight of the victims
(D) Wilson has been involved in a scandal in the city
(E) Wilson has lied about his age

Section 2

8. A certain sprinkler releases water at the rate of 150 liters per hour. If the sprinkler operates for 80 minutes, how many liters of water will be released?
(A) 170
(B) 200
(C) 225
(D) 230
(E) 250

Questions 15-16 refer to the following information.

CAMP SCHEDULE OF CHORES

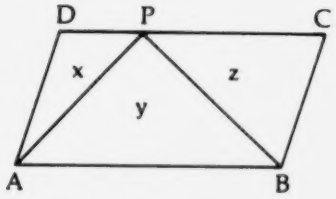
Order of Assignment	Chore
1	Make beds
2	Mop floors
3	Clean windows
4	Pick up litter
5	Empty waste cans
6	Clean bathrooms
7	Pick up mail
8	Inspect cottage
9	Deliver laundry

A boys' camp had 200 empty cottages. When 1,800 boys arrived, they were numbered serially starting with 1 and were assigned, in order, to cottages with 9 boys to a cottage. The first 9 boys were assigned to the 1st cottage, the second 9 to the 2nd cottage, and so on. In each cottage, each boy was assigned to chores according to his number, with the boy having the lowest number in each cottage assigned to the first chore, and so on.

15. What chore will the 994th boy have?
(A) Mop floors
(B) Clean windows
(C) Pick up litter
(D) Clean bathrooms
(E) Deliver laundry

16. What was the number of the boy in the 86th cottage whose assignment was to "inspect cottage"?

(A) 766
(B) 773
(C) 774
(D) 775
(E) 782



19. In parallelogram ABCD above, P represents any point on side DC. If x, y, and z are the areas of the three triangles shown, which of the following CANNOT be the ratio of x to y to z?

(A) 1 to 3 to 4
(B) 7 to 8 to 15
(C) 3 to 7 to 10
(D) 4 to 8 to 12
(E) 2 to 5 to 8

20. If 1/6 is written as a decimal to 200 places, what is the sum of the first 100 digits to the right of the decimal point?

(A) 55
(B) 100
(C) 350
(D) 595
(E) 600

21. A high school basketball team has won 40 percent of its first 15 games. Beginning with the sixteenth game, how many games in a row does the team now have to win in order to have a 55 percent winning record?

(A) 3
(B) 5
(C) 6
(D) 11
(E) 15

22. If $-3 < a < 7$ and if $-2 < b < 0$, which of the following must be true for $(a-b)$?

(A) $-5 < (a-b) < 7$
(B) $-3 < (a-b) < 7$
(C) $-1 < (a-b) < 7$
(D) $-3 < (a-b) < 9$
(E) $-1 < (a-b) < 9$

25. If n is one of three consecutive odd integers, then the possible values of the sum of the 3 integers include which of the following?

I. $3n + 3$
II. $3n$
III. $3n + 6$

(A) I only
(B) II only
(C) III only
(D) I and III
(E) II and III

Section 5

COMPARISON QUESTIONS

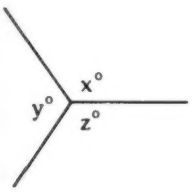
Answer:
A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined
AN E RESPONSE WILL NOT BE SCORED.

Column A	Column B

Two of the three sides of a right triangle R have lengths 7 and 10.

17. Length of the remaining side of R	10

Column A	Column B



$-1 < x < 0$

25. x^2	$-x$

Appendix B: Questionnaire Used with Princeton Review Students (June 1986 SAT)

Survey for Research on SAT Tests

This anonymous questionnaire is designed to help researchers uncover problems students encounter on standardized tests. None of this material will go to your school or be used with your name attached. We need your help—please fill out each question carefully: If you can't answer a question or choose not to, please move on to the next item. Thank you!

1. What is your grade level in school?
   (A) 12th grade or no longer in H.S.
   (B) 11th grade
   (C) 10th grade
   (D) 9th grade or earlier

2. From this list, which is your favorite subject in high school?
   (A) English
   (B) Math
   (C) Social Studies
   (D) Science
   (E) Foreign Language

3. What is your second favorite subject?
   (A) English
   (B) Math
   (C) Social studies
   (D) Science
   (E) Foreign Language

4. How many years of math have you had in high school, from ninth grade until now (include this year, if you are taking math this year)?
   (A) six or more
   (B) five
   (C) four
   (D) three
   (E) two or less

5. How many years of English have you had in high school, from ninth grade until now (include this year, if you are taking English this year)?
   (A) six or more
   (B) five
   (C) four
   (D) three
   (E) two or less

6. How many years of science have you had in high school, from ninth grade until now (include this year, if you are taking science this year)?
   (A) six or more
   (B) five or more
   (C) four
   (D) three
   (E) two or less

7. What is your overall grade average in your high school English courses?
   (A) A to A+ (93–100)
   (B) B+ to A- (87–92)
   (C) B- to B (80–86)
   (D) C to C+ (73–79)
   (E) C- or lower (72 or lower)

8. What is your overall grade average in your high school math courses?
   (A) A to A+ (93–100)
   (B) B+ to A- (87–92)
   (C) B- to B (80–86)
   (D) C to C+ (73–79)
   (E) C- or lower (72 or lower)

9. What is your overall grade average in all your high school courses?
   (A) A to A+ (93–100)
   (B) B+ to A- (87–92)
   (C) B- to B (80–86)
   (D) C to C+ (73–79)
   (E) C- or lower (72 or lower)

10. Thinking of your entire high school class in grade average, are you in the:
   (A) top 5%
   (B) top 10%
   (C) top 25%
   (D) top 50%
   (E) bottom 50%

11. How do you think you compare with other people your own age in your reading and writing ability?
   (A) top 5%
   (B) top 10%
   (C) top 25%
   (D) top 50%
   (E) bottom 50%

12. How do you think you compare with other people your own age in your ability in math?
   (A) top 5%
   (B) top 10%
   (C) top 25%
   (D) top 50%
   (E) bottom 50%

13. Do you feel your past test scores on standardized tests (PSAT, etc.) are accurate?
   (A) No, my ability is higher than the tests indicate.
   (B) Yes, they do reflect my ability.
   (C) No, my ability is lower than the tests indicate.

14. How do you feel about the SAT?
(A) extremely anxious
(B) moderately anxious
(C) somewhat anxious
(D) not anxious at all.

15. Have you taken any other coaching course before this?
(A) yes, in school
(B) yes, outside of school
(C) no

16. Think about the colleges you plan to apply to. Which of these phrases best describes the kind of college that you realistically plan to attend?
(A) academically "super-elite", such as Ivy League, Bryn Mawr, Cal-Tech, Carleton, Chicago, MIT, Stanford, Smith, Swarthmore, Wesleyan, Williams.
(B) academically very strong, such as Bates, Berkeley, Duke, Georgetown, Johns Hopkins, Michigan, Vermont, Virginia, West Point, Wisconsin.
(C) academically strong, such as Fordham, Illinois, North Carolina, Penn State, NYU, Rutgers, SUNY, CUNY, UConn.
(D) academically adequate, such as Monmouth (NJ), CW Post, Sacred Heart (CT), small state colleges, etc.
(E) do not plan to go to a four year college.

17. What is your age?
(A) 18 and over
(B) 17
(C) 16
(D) 15
(E) 14 and under

18. Sex:
(A) Female
(B) Male

19. Ethnic group:
(A) black (Afro-American)
(B) white (not including Hispanic)
(C) Hispanic (Puerto Rican, Cuban, Mexican-American, etc.)
(D) Asian-American
(E) other (including Native American Indian)

20. What is your father's occupation? (Use these categories as accurately as you can. If he is retired, deceased, or not working, answer for his last job.)
(A) lawyer; MD; architect; college professor; manager or owner of medium to large business; high executive in large company
(B) pharmacist; engineer; veterinarian; manager or owner of small business; lower executive in large company; school teacher; pilot; minister
(C) social worker; insurance; real estate salesman; electrician; Armed Forces; foreman; police
(D) carpenter; industrial worker; clerk; sales clerk; truck driver
(E) janitor; carpenter's helper; laborer.

21. What is your mother's occupation? (Use these categories as accurately as you can. If she is retired, deceased, or not working, answer for her last job.)

(A) lawyer; MD; architect; college professor; manager or owner of medium to large business; high executive in large company
(B) pharmacist; engineer; veterinarian; manager or owner of small business; lower executive in large company; school teacher; pilot; minister
(C) social worker; nurse; insurance; real estate salesperson; electrician; Armed Forces; foreman; police
(D) industrial worker; secretary; sales clerk; cashier; maid; nurses aide; waitress; seamstress
(E) housewife; mother; volunteer worker; not in paid job at present.

22. What is your father's education? (If you don't know, answer the best you can).

(A) less than high school graduate
(B) high school graduate
(C) some college
(D) college graduate
(E) graduate or professional (law, M.D., M.A., Ph.D., etc.)

23. What is your mother's education?

(A) less than high school graduate
(B) high school graduate
(C) some college
(D) college graduate
(E) graduate or professional (law, M.D., M.A., Ph.D., etc.)

24. Are you attending a:

(A) public school
(B) parochial school (church-related)
(C) private (prep) school

25. Where is your high school located?

(A) large city (100,000 or more people)
(B) suburb or town in metropolitan area
(C) small city (10,000 to 100,000 people)
(D) rural area or small town (less than 10,000 people, not in metro area)

Thank you again for your help!

Appendix C: Technical Notes for Study of Gender Bias in the June 1986 SAT

by James. W. Loewen

Significance Levels

Tables 2, 3, 5, 8, and 10 are comparisons of percentages based on sample sizes of approximately 500 (all females compared to all males). On such tables, differences of about 8 percent are significant at the .01 level; differences of 6 percent are significant at the .05 level of confidence (two-tailed).

Tables 4, 7, 15, 16, and 17 are comparisons of percentages based on sample sizes of approximately 125 (1/4 of all females, divided into score groups or other groupings, compared to another 1/4, compared to 1/4 of all males, similarly divided, etc.). On such tables, differences of about 13 percent are significant at the .05 level of confidence (two-tailed).

Item "Standardization"

For several years, ETS has been concerned about eliminating what it calls "the contaminating effects of ability differences from the assessment of item fairness." ETS desires to separate out "unexpected differential item performance" from "normal" "differences in subgroup ability." If, for example, we compared sixth-graders to twelfth-graders on the SAT, and sixth-graders did 20 percent worse than twelfth-graders on an item, we would want to know how much worse sixth-graders did on *all* items before concluding that that item was biased against sixth-graders. In ETS's terms, we should compare the two groups using some method that does not "exhibit undesirable sensitivities to differences in overall subpopulation ability" (Dorans and Kulick, 1983, pp. 1-3). We will see that ETS simply uses test score as its measure of "overall subpopulation ability."

In recent years ETS has used several statistical techniques to deal with this problem, including the Mantel-Haenzel technique, transformed item difficulty analysis, and a technique it calls "standardization." Standardization has the advantage of being intuitively clear, and ETS seems to be settling upon it as its method of choice. As ETS researchers Dorans and Kulick put it (1983, Abstract), "the primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items."

"Standardization" as used by ETS does not mean what statisticians mean by the term; hence we will use quotation marks around the term when using ETS's definition. Dorans and Kulick use female/male differences to illustrate the technique; we will follow their example, using item #44 from the Verbal SAT we analyzed, "mercenary is to soldier."

On this item, 48.6 percent of the girls answered correctly compared to 64.3 percent of the boys. Dorans and Kulick would not use that 15.7 percent difference, however, but would "standardize" by overall scores. To do this, they subtract the percent correct among boys who scored 200 on the Verbal SAT from the percent correct among girls who scored 200 on the Verbal SAT; then they do the same for boys and girls who scored 210, and so on, up to those whose overall Verbal SAT score was 800. Then they sum these 61 differences, weighting them by the number of girls in each score category, to calculate d_t, the "standardized" difference.

In practice, this usually results in a percentage difference between the groups which roughly equals the difference between all girls and all boys with which we began, when the two groups have similar overall means. But when the two groups have different means, then "standardization" yields a percentage difference which usually roughly equals the original percentage difference on the item minus the difference in the overall means.[1]

For easier calculation in our example, we grouped our students into 4 "ability" groups rather than 61 and computed d_f, which yielded -15.7 percent, roughly identical to the raw difference. d_f for other verbal items was similar to the raw differences, as Table 1 shows. This was expected, since the girls in this sample scored only .2 worse than the boys overall on the Verbal SAT.

[redacted]

TABLE 1

Raw and "Standardized" Differences on 7 SAT Verbal Items Favoring One Sex by Approximately > 10%

Section, Item No., Description	Female %-Male %	d_f
1 No. 1, "setback," opposite "improvement"	-10.7%	-10.8%
1 No. 5, "sheen," opposite "dull finish"	+18.3	+21.4
1 No. 23, author's tone, science passage	-11.8	-11.7
1 No. 44, "mercenary is to soldier"	-15.7	-15.7
4 No. 21, "pendant is to jewelry"	+ 9.6	+10.0
4 No. 24, "love is to requite"	+14.5	+14.7
4 No. 31, "betrayal" (in human relations item)	+10.2	+10.0

On the math test, "standardization" made a larger difference, again as we would expect, since the boys outscored the girls by 3.5 raw points overall. Table 2 compares the raw and "standardized" differences on each item with > 10 percent differences.

By way of contrast, consider the only item on this Math SAT with any female verbal content, No. 11 from section 2, which includes the name "Judy." Boys outperformed girls on this item by 0.5 percent, making it a relatively good item for girls: when "standardization" is applied, the difference is +5.2 percent, "favoring" girls. A researcher who used "standardized" differences of > 5 percent as the criterion to delete items from this Math SAT would delete "Judy", while leaving five items on the exam that favor boys by more than 10 percent.

A terminology problem afflicts ETS's discussions of "standardization." To compare groups matched in "ability" (or in experience, level of schooling, or the like) appears reasonable. Good researchers would not normally compare apples and oranges, or sixth-graders with twelfth-graders. But overall test score is a circular measure of "ability." Consider this passage by Dorans and Kulick: "Standardization with respect to ability level . . . produces a simple total group comparison, like that based on the overall performance column, which is not confounded by differences in group ability. Standardization accomplishes this goal by using the same standard ability distribution for both groups." (1983, p. 4). A paraphrase could read: "Standardization by total scores produces a simple total group comparison, like that based on the overall performance column, but with the overall group difference removed." The difference is instructive, because ETS's wording can lure its own researchers into imagining that "standardization" is more scientific.

On the contrary, "standardization" can lead to bizarre and paradoxical results. A study of sex differences on the California Achievement Test provides an example

TABLE 2

Raw and "Standardized" Differences on 10 SAT Math Items Favoring One Sex by > 10%

Section, Item No., Description	Female %-Male %	d_f
2 No. 8, "liters per hour"	-10.3%	-5.7%
2 No. 15, "chore 994th boy will have at boys camp"	-12.3	-5.5
2 No. 16, "number of boy with chore at boys camp"	-15.6	-10.9
2 No. 19, "parallelogram ratios"	-12.2	-5.0
2 No. 20, "1/6 as decimal, sum of digits"	-10.7	-2.2
2 No. 21, "basketball team won/loss record"	-27.0	-18.4
2 No. 22, "<(a-b)<"	-11.0	-4.7
2 No. 25, "n as odd integer"	-10.8	-2.9
5 No. 17, "length of right triangle"	-10.7	-3.5
5 No. 25, "inequalities with x^2, -x"	-10.6	-2.3

(Green, 1987).[2] Of the 72 different forms of the CAT examined, girls outscored boys on 69. Looking at simple percentage differences, girls outscored boys by > 5 percent on 1,233 of the 3,102 items, while **not one item** favored boys by > 5 percent. But when "standardization" was applied, only 298 of the 3,102 items showed differences of > 5 percent, and most of those items "favored" boys. In other words, if a sample of girls exceeded boys by 12 percent overall, yet on a given item girls exceeded boys by 6 percent, that item would be one of the 1,233 on which girls outscored boys by > 5 percent, but it would also favor boys by > 5 percent after "standardization."

Thus, even when one group performs dramatically worse than another, such as Blacks vs. whites on the SAT, researchers investigating item bias using "standardization" are just as likely to remove items that favor the lower scoring group as items on which they did particularly poorly. Accordingly, "standardization" is not a tool to locate biased items, at least as the term is commonly defined, but instead may mask bias. While "standardization" is an interesting technique and should be used to supplement raw percentage differences, we would suggest examining simple percentage differences, instead.

Regression Analysis

Examination of scatterplots and correlation and regression coefficients provides another way of analyzing and showing item bias. We plotted the percentage of girls who answered correctly as a dependent variable against the percentage of boys who answered each item correctly. Correlations were very high, as we would expect: r = .970 on the Verbal Section, .987 on the Math. Thus, most items lay very close to the regression line. Nonetheless, the items already discussed in Tables 1 and 2 were observable as outliers on the scatterplots.

On the Verbal SAT, the regression equation was $y = 3.29 + (.946)x$.

Theoretically, for an item which 0 percent of males answered correctly, 3.3 percent of females answered correctly, while for an item which 100 percent of males answered correctly, 97.9 percent of females answered correctly (3.29% + 94.6%).

The regression equation on the math scatterplot was y = -12.3 percent + (1.103)x, implying that for an item which 0 percent of males answered correctly, -12.3 percent of females answered correctly, while for an item which 100 percent of males answered correctly, 97.0 percent of females answered correctly. This regression equation restates what we have already observed, that boys outscored girls on the Math SAT.

Additional References

Dorans, Neil, and Kulick, E. 1983. *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach.* Princeton: ETS.

Notes

1. If the difficulty curve is different for one group, then $d_i \neq$ the percentage difference minus the mean difference.

2. Green used a different statistical manipulation but it had the same effect regarding group means.

Appendix D: Items with Extreme Differences by Sex (November 1987 SAT)

Section 1

2. IRK:
   (A) dilate
   (B) inhibit
   (C) reflect
   (D) soothe
   (E) confront

37. According to the passage, all of the following are correct statements about Comet Brooks before 1886 EXCEPT:
   (A) Its orbital velocity was about 8.1 miles per second.
   (B) Its orbit was considerably larger than it was after 1886.
   (C) Its orbital path crossed the paths of planets in the solar system.
   (D) It could not be detected from Earth.
   (E) It was attracted to the Sun by gravity.

Section 4

2. STAMINA:
   (A) lack of skill
   (B) lack of endurance
   (C) lack of purpose
   (D) disinterestedness
   (E) unwillingness

4. SHEEPISH:
   (A) confident
   (B) prejudiced
   (C) curious
   (D) envious
   (E) amusing

25. Although the undefeated visitors——triumphed over their underdog opponents, the game was hardly the——sportswriters had predicted.
   (A) fortunately..upset
   (B) unexpectedly..classic
   (C) finally..rout
   (D) easily..stalemate
   (E) utterly..mismatch

41. DIVIDENDS:STOCKHOLDERS::
(A) investments: corporations
(B) purchases:customers
(C) royalties:authors
(D) taxes:workers
(E) mortgages:homeowners

Section 2

7. If 2/3 of n is 4, then 1/2 of n is
(A) 1/6
(B) 1/3
(C) 4/3
(D) 2
(E) 3

8. Pat made a total of 48 pottery plates and cups. If she made twice as many plates as cups, how many plates did she make?
(A) 32
(B) 24
(C) 18
(D) 16
(E) 8

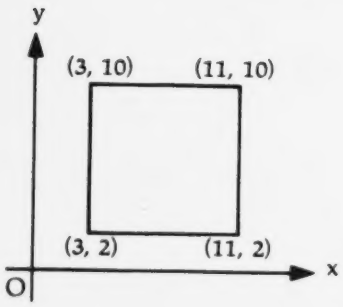
Note: Figure not drawn to scale.

12. In the figure above, if $x = 80$ and $y = 30$, what is the value of k?
(A) 30
(B) 40
(C) 50
(D) 60
(E) It cannot be determined from the information given.

Oatmeal Recipe

Water: 3/4 cup
Salt: 1/4 teaspoon
Oats: 1/3 cup

17. If the least possible multiple of the recipe above is prepared so that a whole number of cups of both water and oats are used, how many teaspoons of salt would be required?
(A) 1/2
(B) 3/4
(C) 1
(D) 2 1/4
(E) 3

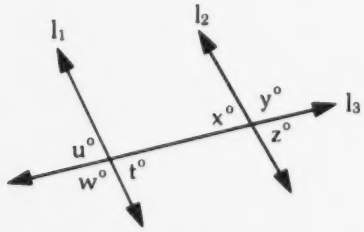


18. In the figure above, a line segment joining the point (3, 10) and which of the following points on the square will separate the square into two regions whose areas are in the ratio of 7 to 1?

(A) (11, 3)
(B) (11, 4)
(C) (11, 6)
(D) (11, 7)
(E) (11, 8)

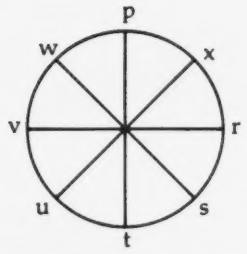
21. Each plant of a certain variety yields 50 seeds in the early fall and then dies. Only 40 percent of these seeds produce plants the following summer and the remainder never produce plants. At this rate, a single plant yielding seeds in 1986 will produce how many plants as descendants in 1989?

(A) 60
(B) 400
(C) 8,000
(D) 16,000
(E) 32,000



23. Lines l_1 and l_2 shown above, are *not* parallel. Which of the following could be true?

(A) u = x
(B) u = y
(C) t = z
(D) w = y
(E) w + x = 180



24. On the circle above, letters opposite each other represent reciprocals; for example, $p = 1/t$. If $pr = x$, which of the following must be true?

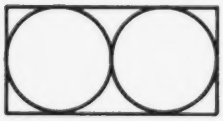
I. $vt = u$
II. $pru = 1$
III. $p + r = v + t$

(A) I only
(B) II only
(C) III only
(D) I and II only
(E) I, II, and III

25. If one of the solutions of the equation $x^2 + x + c = 0$ is 2, what is the other solution?

(A) -3
(B) -2
(C) 0
(D) 3
(E) It cannot be determined from the information given.

Section 5



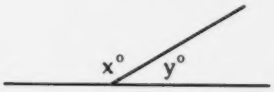
6. The rectangle above contains two circles, tangent to each other and each tangent to three sides of the rectangle. Which of the following pairs of numbers CANNOT be the length and width, respectively, of the rectangle?

(A) 2, 1
(B) 12, 6
(C) 16, 10
(D) 22, 11
(E) 32, 16

Questions 8-27 each consist of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet fill in oval

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined from the information given.
AN E RESPONSE WILL NOT BE SCORED.

EXAMPLES

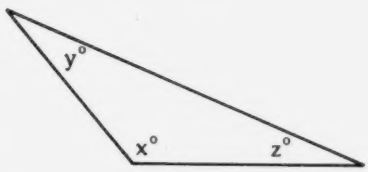
	Column A	Column B	Answers
E1.	2 x 6	2 + 6	● Ⓑ Ⓒ Ⓓ Ⓔ
E2.	 180 - x	y	Ⓐ Ⓑ ● Ⓓ Ⓔ
E3.	p - q	q - p	Ⓐ Ⓑ Ⓒ ● Ⓔ

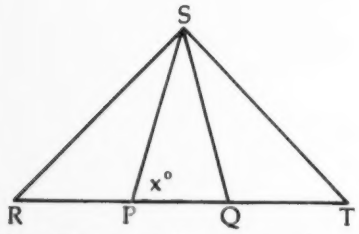
Notes:

1. In certain questions, information concerning one or both of the quantities to be compared is centered above the two columns.

2. In a given question, a symbol that appears in both columns represents the same thing in Column A as it does in Column B.

3. Letters such as x, n, and k stand for real numbers.

Column A	Column B
The gas tank in car R can hold at most 22 gallons of gas.	
10. The number of gallons of gas in car R's gas tank if the tank is 75 percent full of gas	17
 18. The average (arithmetic mean) of x, y, and z	60



Note: Figure not drawn to scale.

29. In the figure above, $\triangle RST$ is a right triangle. $\overline{RS} = \overline{ST}$ and right angle RST has been divided into three equal angles. What is the value of x ?

(A) 65
(B) 70
(C) 75
(D) 80
(E) 85

30. If the price of mints was raised from 5 cents each to 15 cents for 2, what was the increase in price per mint?

(A) 2 1/2¢
(B) 3¢
(C) 5¢
(D) 7 1/2¢
(E) 10¢

31. If a rectangular cake, 9 inches by 13 inches by 2 inches, is cut into x equal rectangular pieces, 3 inches by 3 1/4 inches by 2 inches, and no cake is left over, then x =

(A) 9
(B) 12
(C) 13
(D) 15
(E) 22

33. How many different-sized circles with positive integer radii have areas less than 100?

(A) Four
(B) Five
(C) Six
(D) Ten
(E) Fifteen

35. If s equals 1/2 percent of t, what percent of s is t?

(A) 2%
(B) 200%
(C) 2,000%
(D) 20,000%
(E) 200,000%

The entire November 1987 Form Code 7H may be purchased from The College Board: *5 SATs 1988 Edition*

Item No. 220211
The College Board
Dept. E 83
Post Office Box 6212
Princeton, New Jersey 08541-6212

$6.00 per copy (enclose check payable to The College Board or institutional purchase order).

Appendix E: Distractors Chosen for Questions That Create the Largest Sex Differences (November 1987 SAT)

Percent Chosen for Each of 7 Verbal Questions

QUEST. NO			BOYS			
	A	B	C	D	E	OMITS
Section 1						
2	2.4	4.2	2.9	72.4	5.5	12.6
37	38.1	10.4	6.5	10.9	6.4	27.7
40	1.6	7.2	2.9	20.0	27.5	40.8
Section 4						
2	2.4	90.3	1.1	1.0	3.2	2.0
4	62.9	1.6	8.8	5.1	5.1	16.5
25	9.3	3.1	41.4	17.3	25.3	3.6
41	10.2	8.0	53.2	5.5	13.7	9.5
QUEST. NO.			GIRLS			
Section 1						
2	2.0	3.2	2.2	80.1	4.6	7.9
37	28.7	11.8	7.5	12.4	9.3	30.2
40	2.0	6.6	3.1	23.3	21.1	44.0
Section 4						
2	5.1	78.9	3.2	2.7	6.1	4.0
4	71.3	1.5	7.3	4.1	4.7	11.1
25	13.2	5.8	16.2	39.6	20.6	4.6
41	13.4	11.1	38.9	5.9	19.1	11.6

Percent Chosen for Each of 34 Math Questions

QUEST. NO.			BOYS			
Section 2						
1	1.1	7.6	1.1	1.7	86.9	1.7
7	2.1	2.5	4.6	8.0	75.4	7.3
8	79.5	7.8	1.1	9.0	0.3	2.3
10	2.3	6.8	7.6	7.3	64.4	11.6
12	4.8	4.8	6.3	3.6	72.2	8.2
15	50.3	2.9	40.6	1.9	2.2	2.1
16	2.2	13.7	51.6	2.9	20.2	9.3
17	7.9	16.7	9.5	6.6	33.2	26.1
18	10.8	7.1	7.1	11.5	28.1	35.4
19	6.7	7.6	15.5	13.8	35.2	21.2
20	30.1	20.3	10.7	8.9	12.6	17.4
21	25.3	8.6	23.8	12.2	4.7	25.5
22	5.6	13.2	14.2	25.9	8.4	32.7
23	3.1	29.4	2.8	10.0	35.4	19.3
24	28.2	3.2	9.7	16.5	13.3	29.0
25	15.4	9.4	3.4	2.5	43.7	25.6

Section 5						
6	4.5	1.6	74.7	3.5	1.4	14.5
8	5.0	3.4	90.0	0.7	0.0	0.8
10	7.8	81.2	5.4	2.4	0.0	3.2
14	77.4	9.3	11.7	0.3	0.0	1.3
18	8.8	2.8	69.9	13.8	0.0	4.7
19	14.5	7.1	8.6	56.7	0.0	13.1
20	28.9	7.1	4.6	49.3	0.0	10.1
23	15.0	6.1	42.4	29.8	0.0	6.7
24	45.0	14.1	15.4	15.4	0.0	10.0
26	9.9	25.5	9.2	25.2	0.0	30.2
27	8.8	21.6	33.7	8.3	0.0	27.5
28	5.0	70.3	3.5	5.5	8.2	7.4
29	7.3	3.5	60.1	8.2	3.4	17.4
30	73.9	1.4	10.5	8.2	1.7	4.3
31	9.2	46.4	5.6	5.4	2.7	30.6
32	5.1	9.5	13.8	29.3	13.4	28.9
33	5.7	28.5	8.5	11.3	5.3	40.8
35	18.9	35.6	6.4	7.1	1.2	30.8

QUEST. NO.			GIRLS			
Section 2						
1	1.0	7.9	1.5	2.0	85.6	1.9
7	3.4	4.0	5.6	12.9	62.8	11.4
8	67.4	13.2	1.7	13.7	0.5	3.5
10	2.8	6.7	10.1	8.7	57.1	14.7
12	7.8	5.9	8.9	5.2	61.3	10.8
15	48.1	3.6	38.9	2.7	3.2	3.4
16	2.8	13.5	42.2	4.7	25.1	11.6
17	11.0	19.3	10.9	6.3	23.8	28.7
18	11.5	8.6	7.9	10.3	18.8	42.9
19	8.6	9.6	17.2	13.4	26.1	25.2
20	23.0	19.1	12.8	9.2	15.0	20.8
21	27.9	10.4	14.8	9.3	3.9	33.8
22	6.6	13.5	15.1	19.8	8.3	36.6
23	4.3	20.4	4.1	12.3	37.8	21.2
24	33.8	2.9	11.0	11.9	13.9	26.4
25	9.5	11.0	3.8	2.2	46.6	26.9
Section 5						
6	7.7	2.0	58.9	5.0	2.2	24.0
8	5.0	3.8	89.2	0.9	0.0	1.0
10	11.3	71.7	7.5	4.0	0.0	5.6
14	75.2	12.1	11.0	0.3	0.0	1.3
18	12.8	3.4	57.4	19.5	0.0	6.9
19	18.3	7.0	11.0	47.5	0.0	16.2
20	30.8	8.0	6.8	40.6	0.0	13.8
23	20.1	8.6	39.6	21.9	0.0	9.9
24	36.3	17.8	13.5	18.8	0.0	13.6
26	10.8	18.6	8.7	26.1	0.0	35.7
27	10.6	24.1	24.5	9.6	0.0	31.2
28	5.5	68.9	3.7	5.6	8.7	7.6

29	12.3	4.3	47.4	9.1	3.8	23.2
30	63.4	2.3	13.8	12.7	2.9	4.9
31	10.0	35.2	6.3	7.1	3.5	37.8
32	4.7	10.7	14.1	22.3	14.8	33.4
33	6.3	17.6	9.1	13.0	4.7	49.2
35	26.1	26.6	4.6	3.6	0.9	38.1

TOTAL BOYS= 45391 TOTAL GIRLS= 54606

Appendix F: Questions with Large Percentage Differences for Women of Color Compared to White Women for All Questions (November 1987 SAT)

NO. VERBAL	WHITE	BLACK (B/W)		ASIAN (A/W)		HISPANIC (H/W)		N AMERICAN (N/W)		OTHER (O/W)	
Section 1											
1.	93	75	0.81	80	0.86	80	0.86	87	0.94	85	0.91
2.	84	71	0.85	63	0.75	55	0.65	78	0.93	74	0.88
3.	67	57	0.85	49	0.73	45	0.67	59	0.88	62	0.93
4.	67	53	0.79	55	0.82	50	0.75	58	0.87	61	0.91
5.	64	49	0.77	54	0.84	59	0.92	54	0.84	58	0.91
6.	43	33	0.77	42	0.98	39	0.91	37	0.86	43	1.00
7.	34	26	0.76	30	0.88	24	0.71	29	0.85	31	0.91
8.	23	13	0.57	17	0.74	12	0.52	17	0.74	22	0.96
9.	22	17	0.77	28	1.27	18	0.82	16	0.73	24	1.09
10.	26	17	0.65	24	0.92	19	0.73	21	0.81	24	0.92
11.	96	87	0.91	77	0.80	77	0.80	91	0.95	87	0.91
12.	89	79	0.89	82	0.92	76	0.85	82	0.92	83	0.93
13.	87	73	0.84	76	0.87	78	0.90	79	0.91	79	0.91
14.	60	44	0.73	49	0.82	40	0.67	47	0.78	53	0.88
15.	37	29	0.78	35	0.95	32	0.86	35	0.95	34	0.92
16.	96	91	0.95	83	0.86	82	0.85	92	0.96	89	0.93
17.	87	76	0.87	82	0.94	76	0.87	81	0.93	81	0.93
18.	81	74	0.91	64	0.79	62	0.77	74	0.91	70	0.86
19.	60	54	0.90	48	0.80	46	0.77	54	0.90	54	0.90
20.	56	41	0.73	47	0.84	45	0.80	43	0.77	51	0.91
21.	40	24	0.60	34	0.85	24	0.60	28	0.70	35	0.88
22.	22	12	0.55	19	0.86	13	0.59	18	0.82	20	0.91
23.	14	11	0.79	17	1.21	16	1.14	11	0.79	16	1.14
24.	22	20	0.91	25	1.14	24	1.09	21	0.95	22	1.00
25.	18	15	0.83	18	1.00	14	0.78	15	0.83	20	1.11
26.	17	14	0.82	19	1.12	14	0.82	14	0.82	18	1.06
27.	82	77	0.94	78	0.95	78	0.95	76	0.93	78	0.95
28.	82	69	0.84	73	0.89	72	0.88	73	0.89	75	0.91

29.	38	29	0.76	34	0.89	33	0.87	30	0.79	38	1.00
30.	54	40	0.74	48	0.89	42	0.78	40	0.74	49	0.91
31.	58	48	0.83	54	0.93	49	0.84	54	0.93	56	0.97
32.	53	38	0.72	49	0.92	39	0.74	42	0.79	48	0.91
33.	29	20	0.69	28	0.97	27	0.93	27	0.93	28	0.97
34.	46	34	0.74	45	0.98	36	0.78	37	0.80	43	0.93
35.	29	21	0.72	28	0.97	20	0.69	25	0.86	26	0.90
36.	42	30	0.71	37	0.88	32	0.76	37	0.88	34	0.81
37.	31	17	0.55	26	0.84	18	0.58	22	0.71	24	0.77
38.	49	28	0.57	39	0.80	31	0.63	41	0.84	38	0.78
39.	34	19	0.56	30	0.88	22	0.65	26	0.76	27	0.79
40.	23	12	0.52	18	0.78	13	0.57	15	0.65	16	0.70

Section 4

1.	96	91	0.95	81	0.84	78	0.81	94	0.98	89	0.93
2.	82	62	0.76	70	0.85	56	0.68	73	0.89	74	0.90
3.	86	75	0.87	72	0.84	71	0.83	83	0.97	80	0.93
4.	75	53	0.71	63	0.84	53	0.71	67	0.89	66	0.88
5.	71	58	0.82	61	0.86	53	0.75	64	0.90	63	0.89
6.	70	60	0.86	60	0.86	64	0.91	57	0.81	63	0.90
7.	43	34	0.79	42	0.98	38	0.88	39	0.91	42	0.96
8.	46	35	0.76	36	0.78	30	0.65	38	0.83	45	0.98
9.	47	39	0.83	40	0.85	52	1.11	43	0.91	43	0.91
10.	42	34	0.81	40	0.95	55	1.31	36	0.86	40	0.95
11.	33	23	0.70	27	0.82	27	0.82	28	0.85	30	0.91
12.	37	23	0.62	35	0.95	28	0.76	25	0.68	36	0.97
13.	35	31	0.89	41	1.17	33	0.94	28	0.80	35	1.00
14	22	15	0.68	21	0.95	16	0.73	17	0.77	20	0.91
15.	11	8	0.73	14	1.27	10	0.91	8	0.73	12	1.09
16.	94	85	0.90	82	0.87	85	0.90	91	0.97	88	0.94
17.	87	69	0.79	75	0.86	76	0.87	78	0.90	80	0.94
18.	77	66	0.86	71	0.92	70	0.91	71	0.92	71	0.92
19.	74	61	0.82	67	0.91	65	0.88	65	0.88	68	0.92
20.	66	54	0.82	60	0.91	57	0.86	59	0.89	61	0.92
21.	66	56	0.85	64	0.97	57	0.86	57	0.86	62	0.94

22.	60	56	0.93	58	0.97	54	0.90	56	0.93	58	0.97
23.	42	29	0.69	33	0.79	28	0.67	30	0.71	37	0.88
24.	31	23	0.74	30	0.97	29	0.94	27	0.87	31	1.00
25.	16	11	0.69	17	1.06	15	0.94	10	0.63	18	1.13
26.	82	67	0.82	75	0.91	70	0.85	74	0.90	74	0.90
27.	23	12	0.52	21	0.91	19	0.83	14	0.61	23	1.00
28.	22	11	0.50	21	0.95	17	0.77	13	0.59	22	1.00
29.	32	23	0.72	30	0.94	27	0.84	28	0.88	31	0.97
30.	80	65	0.81	73	0.91	70	0.88	72	0.90	72	0.90
31.	50	36	0.72	46	0.92	41	0.82	41	0.82	43	0.86
32.	21	19	0.90	22	1.05	22	1.05	21	1.00	23	1.10
33.	48	32	0.67	48	1.00	37	0.77	37	0.77	43	0.90
34.	28	20	0.71	25	0.89	24	0.86	19	0.68	24	0.86
35.	14	12	0.86	13	0.93	13	0.93	11	0.79	14	1.00
36.	91	78	0.86	81	0.89	79	0.87	84	0.92	84	0.92
37.	87	75	0.86	80	0.92	77	0.89	80	0.92	80	0.92
38.	77	62	0.81	69	0.90	66	0.86	67	0.87	70	0.91
39.	74	56	0.76	64	0.86	59	0.80	65	0.88	66	0.89
40.	60	41	0.68	49	0.82	44	0.73	53	0.88	53	0.88
41.	42	25	0.60	27	0.64	21	0.50	26	0.62	33	0.79
42.	33	14	0.42	24	0.73	20	0.61	26	0.79	28	0.85
43.	22	16	0.73	22	1.00	15	0.68	18	0.82	21	0.95
44.	19	15	0.79	18	0.95	15	0.79	20	1.05	18	0.95
45.	17	13	0.76	16	0.94	17	1.00	14	0.82	17	1.00

MATH

Section 2

1.	86	78	0.91	87	1.01	81	0.94	79	0.92	82	0.95
2.	96	81	0.84	93	0.97	86	0.90	93	0.97	91	0.95
3.	84	68	0.81	87	1.04	71	0.85	74	0.88	77	0.92
4.	86	70	0.81	86	1.00	77	0.90	79	0.92	80	0.93
5.	90	79	0.88	85	0.94	80	0.89	84	0.93	84	0.93
6.	76	59	0.78	84	1.11	63	0.83	69	0.91	71	0.93
7.	65	43	0.66	75	1.15	49	0.75	51	0.78	58	0.89
8.	70	49	0.70	68	0.97	52	0.74	59	0.84	63	0.90
9.	69	54	0.78	73	1.06	57	0.83	56	0.81	60	0.87
10.	59	41	0.69	62	1.05	46	0.78	54	0.92	54	0.92

11.	68	41	0.60	65	0.96	46	0.68	55	0.81	59	0.87
12.	62	51	0.82	67	1.08	57	0.92	58	0.94	60	0.97
13.	72	52	0.72	74	1.03	56	0.78	61	0.85	63	0.88
14.	41	29	0.71	43	1.05	31	0.76	31	0.76	37	0.90
15.	50	33	0.66	48	0.96	38	0.76	38	0.76	44	0.88
16.	44	25	0.57	52	1.18	31	0.70	31	0.70	40	0.91
17.	25	13	0.52	27	1.08	14	0.56	18	0.72	21	0.84
18.	19	9	0.47	22	1.16	11	0.58	15	0.79	18	0.95
19.	27	16	0.59	33	1.22	17	0.63	23	0.85	25	0.93
20.	23	17	0.74	29	1.26	20	0.87	20	0.87	23	1.00
21.	14	13	0.93	16	1.14	13	0.93	12	0.86	14	1.00
22.	20	14	0.70	26	1.30	15	0.75	14	0.70	18	0.90
23.	21	14	0.67	21	1.00	15	0.71	17	0.81	19	0.90
24.	12	9	0.75	17	1.42	9	0.75	9	0.75	11	0.92
25.	9	4	0.44	17	1.89	6	0.67	6	0.67	8	0.89

Section 5

1.	92	81	0.88	93	1.01	84	0.91	89	0.97	87	0.95
2.	82	67	0.82	85	1.04	71	0.87	70	0.85	73	0.89
3.	79	67	0.85	87	1.10	70	0.89	72	0.91	74	0.94
4.	75	61	0.81	73	0.97	61	0.81	66	0.88	69	0.92
5.	74	57	0.77	83	1.12	61	0.82	62	0.84	67	0.91
6.	61	38	0.62	65	1.07	45	0.74	54	0.89	56	0.92
7.	48	30	0.63	50	1.04	32	0.67	34	0.71	41	0.85
8.	90	82	0.91	91	1.01	83	0.92	84	0.93	86	0.96
9.	80	65	0.81	83	1.04	65	0.81	71	0.89	74	0.93
10.	74	54	0.73	77	1.04	56	0.76	63	0.85	65	0.88
11.	79	68	0.86	85	1.08	66	0.84	69	0.87	75	0.95
12.	57	41	0.72	64	1.12	47	0.82	46	0.81	55	0.96
13.	80	60	0.75	79	0.99	63	0.79	69	0.86	73	0.91
14.	76	65	0.86	79	1.04	66	0.87	70	0.92	71	0.93
15.	58	39	0.67	62	1.07	43	0.74	47	0.81	52	0.90
16.	62	44	0.71	62	1.00	49	0.79	53	0.85	59	0.95
17.	72	58	0.81	79	1.10	58	0.81	58	0.81	66	0.92

18.	59	42	0.71	68	1.15	42	0.71	45	0.76	55	0.93
19.	50	31	0.62	50	1.00	34	0.68	35	0.70	45	0.90
20.	41	32	0.78	51	1.24	34	0.83	35	0.85	40	0.98
21.	63	46	0.73	72	1.14	50	0.79	50	0.79	59	0.94
22.	57	42	0.74	65	1.14	43	0.75	45	0.79	52	0.91
23.	21	18	0.86	31	1.48	19	0.90	18	0.86	21	1.00
24.	37	24	0.65	52	1.41	27	0.73	25	0.68	32	0.86
25.	10	9	0.90	16	1.60	9	0.90	9	0.90	11	1.10
26.	18	15	0.83	26	1.44	16	0.89	14	0.78	19	1.06
27.	25	13	0.52	33	1.32	15	0.60	14	0.56	24	0.96
28.	71	55	0.77	72	1.01	56	0.79	62	0.87	63	0.89
29.	50	25	0.50	60	1.20	36	0.72	36	0.72	43	0.86
30.	67	41	0.61	63	0.94	50	0.75	61	0.91	57	0.85
31.	37	20	0.54	44	1.19	23	0.62	30	0.81	31	0.84
32.	22	16	0.73	30	1.36	18	0.82	16	0.73	20	0.91
33.	18	12	0.67	24	1.33	13	0.72	12	0.67	17	0.94
34.	14	9	0.64	20	1.43	10	0.71	9	0.64	11	0.79
35.	3	2	0.67	7	2.33	3	1.00	3	1.00	4	1.33
TOTALS	40846	4441		2724		2373		601		3621	

Appendix G: Number of Omissions for Each Question, by Sex (November 1987 SAT)

VERBAL QUEST. NO.	GIRLS	BOYS
SECTION 1		
1	465	196
2	4333	5724
3	4212	3755
4	7734	6713
5	7565	7108
6	16174	13793
7	9201	7487
8	2953	2874
9	15355	14672
10	18504	16644
11	330	265
12	346	426
13	640	519
14	1554	1410
15	6427	6510
16	400	462
17	487	460
18	2165	1626
19	1322	1525
20	11876	10564
21	5351	3341
22	13680	10370
23	23328	19460
24	19854	17027
25	24554	21554
26	2672	2525
27	1554	1765
28	1643	1941
29	2046	1945
30	2319	2142
31	2251	2147
32	4524	4203
33	4531	4315
34	6291	6082
35	9778	8861
36	13124	10209
37	16497	12594
38	18012	13910
39	23170	17722
40	24007	18509

SECTION 4

1	678	460
2	2211	914
3	1434	957
4	6066	7500
5	4129	4522
6	4374	3798
7	9454	8090
8	8775	7007
9	11371	9229
10	13449	12304
11	18235	15088
12	14192	13798
13	16511	15351
14	14098	14704
15	11099	10265
16	503	321
17	1087	1137
18	918	742
19	4683	3632
20	4084	2865
21	2465	2234
22	4019	3963
23	3870	2622
24	9120	7122
25	2519	1654
26	4372	4015
27	3701	3749
28	7887	6744
29	9450	8545
30	5170	5118
31	8473	7729
32	9693	9118
33	9271	8772
34	16922	14409
35	18210	16222
36	954	991
37	1364	1455
38	5574	4571
39	3123	2563
40	2843	3156
41	6361	4290
42	4767	3703
43	8780	6443
44	25052	18294
45	27091	22137

	MATH QUEST. NO.	GIRLS	BOYS
SECTION 2	1	1050	752
	2	162	59
	3	1671	1241
	4	409	206
	5	319	213
	6	5213	3599
	7	6199	3310
	8	1900	1036
	9	8471	5392
	10	8001	5282
	11	5873	4035
	12	5894	3730
	13	4633	3668
	14	13257	10226
	15	1879	962
	16	6349	4225
	17	15697	11836
	18	23415	16073
	19	13747	9609
	20	11368	7908
	21	18448	11578
	22	20012	14842
	23	11567	8747
	24	14432	13146
	25	14697	11641
SECTION 5			
	1	124	59
	2	3326	2038
	3	4696	3350
	4	2526	1631
	5	3602	2418
	6	13117	6565
	7	8557	7141
	8	559	378
	9	2160	1188
	10	3068	1455
	11	2821	1742
	12	2845	1838
	13	2962	1592
	14	737	588
	15	968	599
	16	3899	2054

17	5724	3902
18	3745	2149
19	8845	5948
20	7528	4564
21	3980	2298
22	4321	3420
23	5382	3044
24	7445	4535
25	16346	11953
26	19521	13701
27	17053	12465
28	4126	3381
29	12667	7903
30	2665	1945
31	20641	13905
32	18255	13119
33	26884	18501
34	22987	17820
35	20820	13971

TOTAL OMITS GIRLS, VERBAL = 679631 BOYS, VERBAL = 585658
TOTAL OMITS GIRLS, MATH = 499565 BOYS, MATH = 346478

Appendix H: Females' Average Scores Are Lower Than Males' At Each Income Level

1988 National Sex/Ethnic Profiles
College Entrance Examination Board

SAT Scores By Family Income

1988 Profiles for Males	Number of SAT Takers	Percent	SAT-V Mean	SAT-M Mean	Combined Score
Income					
Less than 10,000	19,442	4	372	450	822
10,000-20,000	52,936	11	400	465	865
20,000-30,000	78,175	17	422	483	905
30,000-40,000	93,073	20	434	494	928
40,000-50,000	67,094	14	446	508	954
50,000-60,000	49,880	11	455	518	973
60,000-70,000	30,212	6	461	524	985
70,000 or more	74,494	16	474	542	1016
No response	78,759				

1988 Profiles for Females	Number of SAT Takers	Percent	SAT-V Mean	SAT-M Mean	Combined Score
Income					
Less than 10,000	28,835	6	356	397	753
10,000-20,000	69,798	14	387	421	808
20,000-30,000	88,877	18	410	440	850
30,000-40,000	100,105	20	423	453	876
40,000-50,000	68,885	14	436	467	903
50,000-60,000	50,240	10	444	477	921
60,000-70,000	30,017	6	451	485	936
70,000 or more	69,614	14	463	502	965
No response	83,928				

FEMALS'AVERAGE SCORES ARE LOWER THAN MALES AT EACH EDUCATIONAL LEVEL

1988 National Sex/Ethnic Profiles
College Entrance Examination Board

SAT Scores by Family Educational Level

1988 Profiles for Males

Highest Level of Parental Education	Number of SAT Takers	Percent	SAT-V Mean	SAT-M Mean	Combined Score
No High School Diploma	18,575	4	355	438	793
High School Diploma	174,450	35	408	469	877
Associate Degree	33,983	7	419	478	897
Bachelor's Degree	141,250	29	452	517	969
Graduate Degree	125,490	25	482	546	1028

1988 Profiles
for Females

Highest Level of Parental Education	Number of SAT Takers	Percent	SAT-V Mean	SAT-M Mean	Combined Score
No High School Diploma	25,398	5	341	389	730
High School Diploma	212,647	39	397	428	825
Associate Degree	38,224	7	409	439	848
Bachelor's Degree	144,555	26	441	475	916
Graduate Degree	127,813	23	470	502	972

Appendix I: Opinion and Order By Judge John M. Walker

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

KHADIJAH SHARIF, by her mother and next friend, AMIDA SALAHUDDIN, *et al.*, individually and on behalf of all others similarly situated,

Plaintiffs,

—against—

NEW YORK STATE EDUCATION DEPARTMENT; and THOMAS SOBOL, Commissioner of Education, in his official capacity,

Defendants.

88 Civ. 8435 (JMW)

OPINION AND ORDER

WALKER, District Judge:

This case raises the important question of whether New York State denies female students an equal opportunity to receive prestigious state merit scholarships by its sole reliance upon the Scholastic Aptitude Test ("SAT") to determine eligibility. To the Court's knowledge, this is the first case where female students are seeking to use the federal civil rights statute prohibiting sex discrimination in federally-funded educational programs to challenge a state's reliance on standardized tests. This case also presents a legal issue of first impression: whether discrimination under Title IX can be established by proof of disparate impact without proof of intent to discriminate.

After careful consideration, this Court finds that defendants are discriminating against female plaintiffs and their putative class in violation of Title IX and the equal protection clause of the U.S. Constitution. For the reasons set forth below, this Court enjoins the State Education Department and its Commissioner from awarding the merit scholarships at issue solely on the basis of the SAT.

I. The Present Action

In November, 1988, plaintiffs—ten high school students, individually and on behalf of all others similarly situated, and two organizational plaintiffs[1]—brought an action for declaratory and injunctive relief against the State Education Department ("SED") and Commissioner of Education Thomas Sobol, in his official capacity, alleging that New York's exclusive reliance on the SAT to award Empire State and Regents scholarships discriminates against female students in violation of the equal protection clause of the Fourteenth Amendment to the U.S. Constitution, Title IX of the Education Amendments of 1972, 20 U.S.C. §§ 1681 *et seq.*, as amended by the Civil Rights Restoration Act of 1987, Pub. L. 100-265, and the regulations pursuant to Title IX, 34 CFR Part 106. Plaintiff's proposed class is composed of "all

female high school seniors in New York State who are or will be applicants for Regents' College Scholarships and Empire State Scholarships of Excellence." Am. Complaint at para. 4.[2]

In essence, plaintiffs contend that the SED's reliance upon the SAT disproportionately impacts female students without advancing the legislature's purpose of recognizing and awarding superior high school achievement. Plaintiffs argue: "(1) the SAT was not designed to measure academic performance and achievement, and cannot appropriately be put to that use, (2) but even if it did, the SAT discriminates against female applicants for scholarships, because it underpredicts academic performance for females as compared to males." P. Mem. at 5.

On December 21, 1988, plaintiffs filed an order to show cause as to why this Court should not issue a preliminary injunction enjoining SED's practice of exclusive reliance on SAT scores in awarding Regents and Empire State scholarships. On that date, in a conference before this Court, defendants represented that, to cover the possibility of an adverse decision that would require the use of grade point averages (variously "GPAs") to determine scholarship eligibility, the SED would commence collection of GPAs immediately.

On January 12, 1989, defendants submitted a cross-motion for an order dismissing the complaint on the grounds that the Court lacks subject matter jurisdiction, that venue is improper, and that the complaint fails to state a claim on which relief can be granted.

On January 23, 1989, at a hearing, the Court accepted *amici* briefs of the Educational Testing Service ("ETS") and the College Entrance Examination Board, and the Hewitt School District, and heard the testimony of educational testing experts, college deans of admission, and SED administrators with knowledge of the SED's programs of scholarship and testing practices. The Court carefully examined the submissions of the parties, assessed the credibility of the witnesses and reviewed word by word the hearing transcript.

II. Background

A. Evolution of New York State Scholarship Awards

New York State, in one of the most extensive merit scholarship programs in the country, each year makes 26,000 academic achievement awards to New York's high school graduates. In order to understand the program's current purpose, a brief recitation of the program's evolution is appropriate.

1. Reliance Upon College Entrance Diplomas and Special Regents Examinations

New York State's scholarship program began in 1913, when the legislature first awarded 750 Regents Scholarships in the amount of $100 a year for a period of four years. Act approved Apr. 16, 1913, ch. 292, 1913 N.Y. Laws, § 527. At that time, the $100 stipend was sufficient to cover the tuition charged at most colleges in the State.[3] Thus, the award was in the nature of a full scholarship which would promote excellence in education by enabling "the most deserving and meritorious students . . . [to] obtain a college or university training, many of whom would be deprived of such education were it not for the wisdom of the State in providing these scholarships."[4]

The 1913 law authorized the State Board of Regents to make all rules governing the award of the scholarships. Ch. 292, 1913 N.Y. Laws § 72. From 1913 until 1944, the State determined scholarship winners based upon the results of general high school Regents examinations, which also were the basis for granting the college entrance diploma. Lott, T. 64.[5]

By 1944, the SED recognized that it could no longer rely solely upon general high school Regents examinations and college entrance diplomas in awarding Regents Scholarships. First, it was hard to rank students based upon the college entrance diploma because it was "difficult under the statute to know just what subjects to take into account in computing the averages of pupils."[6] Second, the nature of the high

school general Regents exams had changed. Instead of measuring levels of achievement in the variety of courses taught in high school, the general Regents exams became a test of the bare minimum that a student needed to know to graduate from high school, and thus was a poor method for sorting students at the top of the spectrum. Lott T. at 73. Faced with these difficulties, in 1944 the SED developed a separate, more challenging Regents scholarship examination. Meno Aff. para. 2. The examination, in use for the next twenty years, was divided into two equal parts—aptitude and achievement—and was six hours long. Lott T. at 68.

In 1974, New York State's scholarship program changed dramatically following a reevaluation by a Select Committee on Higher Education. At the time the Regents scholarship program provided an annual award of $1,000 to a limited number of highly qualified students. The Committee found that the legislature's goal of substantially funding students' college educations as an incentive for select students to attend college was no longer being met.[7] The Committee found that the Regents scholarship examinations "can be criticized for actually rewarding the family background and upbringing that enables students to study and perform well, rather than an objective kind of merit."[8]

Prompted by these concerns, the legislature restructured its awards, creating two types of awards: first, "general awards" which provide substantial monetary assistance, and second, "academic performance awards" which recognize achievement. Act of 1974, ch. 942, N.Y. Laws §§ 604, 605. Classifying Regents Scholarships as "academic performance awards," the legislature reduced the awards to a stipend of $250, and increased the number of awards to 25,000 to be allocated by county of residence. The legislature created additional awards for the least competitive high schools to enable them to receive at least one for every forty graduates from that school in the preceding year. Ch. 942, 1974 N.Y. Laws § 605 (1) (b).

In the "general awards" category, the legislature created a Tuition Assistance Program ("TAP") to fund college students based upon financial need. The legislature made TAP awards available to all students enrolled in approved programs and are given to those who demonstrate the ability to complete such program's courses, and who satisfy financial need requirements established by the Commissioner of Education. Ch. 942, 1974 N.Y. Laws § 604; N.Y. Educ. Law § 667(1) and (4) (McKinney 1988 & Supp. 1989).

2. Reliance Upon the SAT

In 1977, as a cost-cutting measure, the legislature eliminated its funding for the Regents scholarship examinations. P. App. I, Ex. 3. Instead, the legislature directed that the scholarships be awarded on the basis of "nationally established competitive examinations." Act approved Apr. 12, 1977, ch. 63, 1977 N.Y. Laws § 1. The SED considered examinations, including: (1) the SAT; (2) the American College Testing Program ("ACT"), (3) the Preliminary Scholastic Aptitude Test ("PSAT"); and (4) a combination of Achievement Tests, individual tests given in particularized areas of study, including biology, chemistry and foreign languages. Lott T. at 66. While the Achievement Tests measured performance in a wide variety of courses in a high school curriculum, the SED did not want to require them to do so for Regents scholarship purposes.[9] The SED similarly rejected the ACT, a test much like the SAT, because few students take the ACT in New York.[10] The PSAT, a shorter version of the SAT given in the junior year of high school, was not considered a viable option beause students took it too early in their high school careers.

By process of elimination, then, the SED chose the SAT, the test taken by the greatest number of students.[11] Unlike the Regents scholarship examinations, the entire SAT is labeled an "aptitude" test, and the SAT only purports to test two subjects—Math and English. D. Mem. at 8. Despite the SED's claim that the Regents Scholarship exam and the SAT are very similar, D. Mem. at 9, the State's own witness, Lynn Richbart, [12] testified that about 30 percent of the SAT questions would not have appeared on the Regents Scholarship exam. Richbart T. at 184. Richbart testified that SAT questions, unlike Regents Scholarship questions, often

require that students remember concepts learned in earlier grade levels, or test students on material which is outside the high school curriculum. Richbart T. at 178-182.[13] Moreover, the SAT requires students to be able to answer questions designed specifically for the SAT, such as comparison and logic questions, that are not present on the Regents Scholarship exams. Richbart T. at 182. In addition, the SAT was never designed to test high school achievement. While high school performance may affect a student's performance on the SAT, the SAT does not cover the high school curriculum—indeed there is no standard high school curriculum in New York State—nor has it ever been validated to test achievement in high school. *See* validation discussion *infra*.

In 1986, the legislature created the more selective Empire State Scholarships of Excellence carrying an annual stipend of $2,000 to be awarded to the 1,000 highest ranking Regents Scholarship winners. Act approved and effective Apr. 18, 1986, ch. 56, 1986 N.Y. Laws. The Governor's approval memorandum to the legislative bill creating the Empire State Scholarships stated the purpose of these new awards as follows:

> The *Empire State Scholarship of Excellence Program will recognize academic achievement* and provide a significant inducement for New York's brightest students to continue their studies in the State. These new scholarships will complement the eforts we are undertaking to acknowledge and enhance the *educational performance* of our brightest youth. (emphasis added).

1986 Legislative Annual, P. App. I, Ex. 3. Like the Regents Scholarship, the Empire State Scholarship is distributed by county of residence, and is renewable for five years. There is no minimum quota per high school for Empire State Scholarships.

3. Reliance Upon SATs and GPAs: The 1987 Experiment

In response to allegations that the SED's practice of relying solely upon the SAT in awarding Regents and Empire State Scholarships discriminated against females who consistently scored below males, the Board of Regents asked the Governor and legislature for $100,000 to develop a new scholarship achievement examination. The legislature declined to fund a special examination but, instead, amended the Education Law to require that the awards be based in part upon the student's grade point average ("GPA") as a measure of high school achievement. Senator Kenneth Lavalle, introducing the legislation, explained that the "statute intended to correct a gross inequity that pervaded the New York educational system caused by awarding of Regents College Scholarships and Empire State Scholarships of Excellence based solely on the results of a nationally administered standardized examination.[14] Lavalle Aff. para. 2, P. Reply Mem., Ex 7. The SED specified in its announcement of the new legislation to high school principals that the law was changed "[i]n order to provide for a better balance of male and female winners." P. App. I, Ex. 10.

The new legislation, for the first time, expressly stated that awards are to be based on a measure of "high school performance." Act approved and effective Aug. 7, 1987, ch. 837, 1987 N.Y. Laws §§ 1, 2. In doing so, the legislature altered the criteria for scholarship eligibility—on a one-year, experimental basis—to require the SED Commissioner to base awards on a formula which at least includes a measure of high school performance, and which may include nationally established competitive examinations. The amendment also required the Commissioner to "complete a statistical review of the gender, racial and ethnic composition of students awarded such scholarships within sixty days of the announcement of such scholarship award." *Id.* The legislation included a sunset provision that provided that the amendment would automatically lapse after one year if it were not affirmatively extended.

In May 1987, the SED examined possible measures of high school performance that could be used to select scholarship winners equitably. May Hearing at 9. The SED surveyed high school principals for information concerning grade point averages and class rank. *Id.* at 4. The possibility of using class rank as a measure of

high school performance was dropped for three reasons: (1) it is not used by all schools; (2) it adversely affects students in highly selective schools; (3) it cannot be used to compare students from schools of different size. *Id.* at 19; Lott T. at 72.

The SED also found drawbacks to the use of grade point averages. Because of the volume of scholarship applications it receives yearly, approximately 100,000, the SED would be unable to individually evaluate the GPA information submitted for each candidate as is done by college admission committees. Byrne aff. § 3; Sharrow T. at 133. Also, the SED concluded that it was difficult to convert grade point average information to a common scale because: (1) there is a lack of comparability in the substance of the courses for which grades are given; (2) school grading practices differ from district to district and different grading scales are utilized; (3) schools differ in their practice and philosophy regarding weighting grades in order to take into account course difficulty; and (4) reported grades may reflect grade inflation. Meno Aff. para. 2-5; Sharrow T. at 129; Meno T. at 140. On the other hand, the SED's survey indicated that there is a great deal of uniformity as to grading scales: 85 percent of the public schools and 73 percent of the private schools used a numerical score of 1-100. Results of High School Survey, P. App. I, Ex. 8.

Despite comparability difficulties, the SED chose to use GPAs as the best available measure of high school achievement. In awarding the Regents and Empire State Scholarships for the 1988 graduates, the SED gave equal weight to students' SAT scores and GPAs, as the measure of high school performance. The SED, however, did not issue specific instructions to schools as to how grades should be reported.[15] As a result, some schools reported weighted grades, taking into account course difficulty, while others reported students' grades as they appeared on their transcripts. Hamburger T. at 3-4. Such inconsistent reporting practices touched off a controversy among school administrators who accused each other of cheating in weighting and reporting grades. Meno. T. at 139.

In 1988, under the procedure using a combination of grades and SATs weighted equally women received substantially more Regents and Empire State Scholarships than in all prior years in which the SAT had been the sole criterion. P. App. I, Ex. 2. In both 1987 and 1988, young women comprised approximately 54 percent of the applicant pool for the scholarship, yet the results in 1988 when grades and SATs were used were markedly different. The results are summarized as follows:

	Winners of Empire State Scholarships of Excellence		Winners of Regents College Scholarships	
	Males	Females	Males	Females
1988	62	38	51	49
1987	72	28	57	43

When GPAs were used in 1988, the legislature held hearings to evaluate the new practice of using both GPAs and SATs. Although use of GPA information reduced the disparity between the number of males and females receiving Scholarships, Commissioner Sobol recommended that the practice be discontinued, as soon as a new scholarship exam was developed, because: (1) use of GPA information put an increased burden on school staff; (2) use of GPA did not provide an equitable way to compare students from different schools; and (3) use of GPA would encourage students to avoid more challenging courses in order to obtain better grades for Scholarship purposes. May Hearing at 17-18. Sobol requested funds for a new scholarship exam but also recommended that, until a separate Regents Scholarship examination could be established, GPAs continue to be used in conjunction with SAT scores. *Id.*[16]

Despite Commissioner Sobol's recommendation, the legislature allowed to lapse the eligibility calculation "based on a formula which includes high school performance and which may include nationally competitive examinations." The

standard thereby reverted to awards "on the basis of nationally established competitive examinations." In the 1988 legislative session, the SED received funds for a new scholarship examination, but has not yet received approval for a developed test. Meno. Aff. para. 9. In September, 1988, the SED determined that it would award Regents and Empire State Scholarships to 1989 high school graduates on the basis of SAT scores alone. It is the SED's sole reliance on SAT scores for 1989 graduates that plaintiffs complain denies them equal protection under the Fourteenth Amendment to the U.S. Constitution and violates Title IX of the Education Amendments of 1972.

B. Use of the SAT for Merit Scholarship Awards

1. ETS Recommendations and States' Practice

The Educational Testing Service ("ETS") developed the SAT in order to predict academic performance in college. Willingham Aff. at paras. 5-6. The ability of the SAT to serve this purpose has been statistically "validated." Willingham Aff. at paras. 16-19.[17] It is undisputed, however, that the SAT predicts the success of students differently for males and females. Willingham Aff. at para. 32. In other words, while the SAT will predict college success as well for males within the universe of males as for females within the universe of females, when predictions are within the combined universe of males and females, the SAT *underpredicts* academic performance of females in their freshman year of college, and *overpredicts* such academic performance for males.[18] The SAT has never been validated as a measure of past high school performance.

Both the ETS and the College Board, which administers the SAT, specifically advise against exclusive reliance upon the SAT, even for the purpose for which the SAT has been validated—predicting future college performance.[19] Instead, ETS researchers recommend that college admissions counselors use a combination of high school grades and test scores because this combination provides the highest median correlation with freshman grades. Tittle Aff. at paras. 25-29. Additionally, the National Association of College Admission Counselors ("NACAC") Code of Ethics requires member institutions to refrain from using minimum test scores as the sole criterion for admission, to use test scores in conjunction with other data such as school record and recommendations, and to refrain from using tests in any manner that may discriminate against students.[20] Thus, many colleges refrain from using test scores exclusively to decide admissions questions. *See* Stewart T. at 58-59; Sharrow T. at 122; Behnke Aff. at paras. 2, 7, 8; Mason Aff. at paras. 5, 6, 7.

Notwithstanding ETS and NACAC guidelines recommending against using the SAT as the sole basis on which to award scholarships or offer admissions, the SED adopted such a policy in 1974. New York State is one of only two states in the nation to rely solely on SAT scores for the award of state sponsored merit scholarships instead of factoring in other measurements, such as grade point average or high school rank. May Hearing at 54. Lee Aff. at para. 3. Most states rely, at least in part, upon GPAs. For instance, California's extensive merit scholarship program, which gives nearly 17,000 awards annually, relies upon self-reported GPAs. Moss Aff. at paras. 1, 3, 5-6.

2. SAT as Measure of High School Performance

Both the Empire State and Regents Scholarships are intended to reward past academic achievement of high school students who have demonstrated such achievement to pursue their educations in New York State. Lott T. at 91; Memo Aff. at para. 9. It is undisputed, however, that the SAT was developed and validated to serve a different purpose—*predicting* performance *in college*.

Professional standards governing educational testing require statistical analysis ("validation") to be undertaken to ensure that a test is properly used for its intended purpose. Shapiro T. at 51. For example, the American Psychological Association's Standards on Psychological Testing require that "evidence of validity should be presented for the major types of inferences for which the use of a test is

recommended." P App. I, Ex. 6, at6 13. Similarly, the College Board requires that tests be validated periodically "to ensure that they predict the expected outcome at a level acceptable for the institution's particular purpose." 1987-1988 ATP Guide for High Schools and Colleges. The SED has never validated the SAT for the purpose of measuring high school performance. Lott. T. at 89.[21]

Notwithstanding the absence of validation studies, it is the SED's current position that the SAT provides a good measure of high school performance because it "measures skills and knowledge primarily developed in school." Byrne Aff. para. 17. The SED does not dispute that the SAT does not measure performance in all high school courses, but claims merely that the SAT partially tracks high school English and math courses and thus tests achievement. Lott T. at 89.[22] The SED concedes that the SAT does not measure achievement in other subject matters such as science, social studies, and foreign languages. Moreover, the SED concedes that overall GPAs are a better measure of high school performance than SATs. Lott. T. at 90; *See also* Anastasi, P. App. II, Part 2, Ex. C.

3. Statistical Impact on Men and Women Statewide

Males have outscored females on the verbal portion of the SAT since 1972, with an average score differential of at least 10 points since 1981. Males have also consistently outscored females on the mathematics portion, with an average differential of at least 40 points since 1967.[23] In 1988, for example, girls scored 56 points lower than boys on the test. The probability that these score differentials happened by chance is approximately about one in a billion and the probability that the result could consistently be so different is essentially zero. *See* Gray Aff. at para. 6.

Statisticians have attempted to explain the score differentials between males and females by removing the effect of "neutral" variables[24], such as ethnicity, socioeducational status (parental education), high school classes, and proposed college major. However, under the most conservative studies presented in evidence, even after removing the effect of these factors, at least a 30 point combined differential remains unexplained.[25]

As a result of the State's practice of basing scholarship awards solely upon SAT scores, males have consistently received substantially more scholarships than females. In 1987, for example, males were 47 percent of the scholarship competitors, but received 72 percent of the Empire State Scholarships and 57 percent of the Regents Scholarships.[26] For Empire State Scholarships, these results represent 15.8 standard deviations from the mean; for Regents Scholarships, the difference represents 31.7 standard deviations. In other words, the probability that the Empire State Scholarship results would occur by chance is less than one in a billion, and the probability that the Regents Scholarships results would occur by chance is even less. Shapiro T. at 29.[27]

III. Discussion

A. Procedural Issues

At the outset, defendants argue that this Court should dismiss plaintiffs' complaint on three procedural grounds: first, this Court is without authority to issue the relief requested in this case because plaintiffs do not have standing to bring their claims; second, the Court lacks subject matter jurisdiction; and third, venue is improper. The Court will consider each of these arguments in turn.

1. Standing

In order to establish standing for the purposes of the constitutional "case or controversy" requirement, the general rule is that a plaintiff "must show that he personally has suffered some actual or threatened injury as a result of the putatively illegal conduct of the defendant," *Gladstone Realtors v. Village of Bellwood*, 442 U.S. 91, 99, (1979), and that the injury is "likely to be redressed by a favorable decision." *Simon v. Eastern Kentucky Welfare Rights Organization*, 426 U.S. 26, 38 (1976). Otherwise, the

exercise of federal jurisdiction "would be gratuitous and thus inconsistent with the Article III limitation." *Id.* at 38.

More precisely, plaintiffs must demonstrate: (1) that the "interest sought to be protected is within the zone of interests protected or regulated by the statute or constitutional guarantee in question," *Association of Data Processing Service Organizations Inc. v. Camp*, 397 U.S. 150, 153 (1970), (2) "injury in fact" and (3) "causation in fact."

Plaintiffs have fulfilled the first standing requirement. The interest sought to be protected—freedom from discrimination in the award of state scholarships—is within the zone of interests to be protected by the Fourteenth Amendment and Title IX, Education Amendments of 1972, 20 U.S.C. § 1681.

The second requirement, injury in fact, is satisfied by a showing of a likelihood of harm, if not actual harm. In *University of California Regents v. Bakke*, 438 U. S. 265, the Supreme Court found that a student had standing to challenge a school's allegedly discriminatory admissions policy, not because he could establish that he would have been admitted were it not for the challenged policy, but rather because his *chances* for admission were reduced by the policy. *Id* at 280-81 n. 14 (Powell, J., concurring). *See also, McCleskey v. Kemp*, 481 U.S. 279, 295 n. (1987); *Heckler v. Matthews*, 465 U.S. 728, 738 (1984); *Gladstone Realtors v. Village of Bellwood*, 441 U.S. at 115.

Plaintiffs here allege that their *chances* for winning a state merit scholarship are reduced by the SED's practice of basing such awards solely on SAT scores and that, therefore, they are less likely to receive benefits such as substantial public recognition, an enhanced ability to attract additional scholarships, and an increased opportunity to attend the college or university of their choice.[28] These allegations alone are sufficient to establish "injury in fact."[29]

Moreover, while plaintiffs need only establish a likelihood of injury, they have shown as to at least three plaintiffs a near certainty of injury if the SED is not enjoined. Defendants concede that plaintiffs Hart, Capodice, and Bozon probably will qualify for Regents Scholarships if eligibility is determined by using equally weighted GPA and SAT scores but will *not* qualify if SAT scores are the sole criterion. T. at 19, 207; Byrne Aff. at para. 10. These three plaintiffs alone are sufficient to establish standing to challenge the awarding practices for both Regents and Empire State Scholarships since both are awarded from the same list of 25,000 names.[30] Because the claims raised by plaintiffs necessarily implicate the entire system, and any relief would require modification of that system, plaintiffs have standing to challenge both the Regents and Empire State Scholarships even though they personally may not be eligible for the latter.

The final requirement, causation in fact, necessitates that the injury be both "fairly traceable" to the defendant and "redressable." *Allen v. Wright*, 468 U.S. 737, 753 n.19 (1984). As with injury in fact, causation in fact does not require a showing of complete certainty. In the Second Circuit:

> All that is required is a showing that such relief be reasonably designed to improve the opportunities of a plaintiff not otherwise disabled to avoid the specific injury alleged. To ask the plaintiffs to show more than that they would 'benefit in a tangible way from the court's intervention,' would be to close our eyes to the uncertainties which shroud human affairs.

Huntington Branch N.A.A.C.P. v. Town of Huntington, 689 F. 2d 391, 394 (2d Cir. 1982) (emphasis added) (plaintiffs seeking funding to construct housing project had standing to challenge zoning ordinance, even though no federal housing money was presently available).

In the present case, plaintiffs allege that the SED's reliance upon the SAT is the direct cause of their injury. Injunctive relief compelling the SED to use an alternative procedure with a less discriminating effect would redress their grievance. Defendants argue that because variables other than sex might account for the disparate number of women receiving low SAT scores—and, consequently, *not* receiving scholarships—there is no causation. D. Mem. at 23. This, however, is a

dispute on the merits of plaintiffs' claim. Standing does not depend on whether plaintiffs actually will prevail. *See e.g., McCleskey* 481 U.S. 279.

The fact that some of the named plaintiffs may not receive scholarships if the injunction is granted presents no barrier to this suit. The claim rests on the alleged discriminatory nature of the system as a whole. In analogous circumstances, the Supreme Court held that a black would-be resident had standing to challenge discriminatory zoning practices, because he intended to apply for housing, although he might not actually obtain it. *Village of Arlington Heights v. Metropolitan Housing Authority*, 429 U.S. 252, 264 (1977). Here, as in *Arlington*, if the requested relief is granted, the plaintiffs would no longer suffer the injury complained of. *See also Pennell v. City of San Jose*, ___U.S.___, 108 S. Ct. 849 (1988) (landlords had standing to challenge rent control ordinance when there was a likelihood of enforcement of the ordinance and concomitant probability that rent would be reduced below what some landlords could afford); *Duke Power Co. v. Carolina Environmental Study Group*, 438 U.S. 59 (1978) (standing to challenge act limiting liability in the event of a nuclear accident where "substantial likelihood" that construction of plants could not be completed without liability limit).

Since we find plaintiffs have sufficiently alleged injury in fact and causation in fact and that plaintiffs are within the requisite zone of interests, we conclude that plaintiffs have standing.

2. Jurisdiction

Defendants argue that this Court lacks subject matter jurisdiction over plaintiffs' equal protection claim against the SED. This argument is wholly meritless.

While it is true that the Eleventh Amendment bars suits against states, *Pennhurst State School & Hospital v. Halderman*, 465 U.S. 89, 100 (1984), the "important exception to this general rule [is that] a suit challenging the constitutionality of a state official's action is not one against the state." *Id.* at 102. Moreover, Congress has specifically provided that a state shall not be immune from suit under Title IX. 42 U.S.C. § 2000d-7. Thus, this Court has jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1343 (3) and (4).

3. Venue

Defendants also challenge the venue of this action. They argue that it is more properly brought in the Northern District of New York where the defendants are located. The Court finds this argument unpersuasive.

Venue in this case is governed by 28 U.S.C. § 1391(b), which provides:

> A civil action wherein jurisdiction is not founded solely on diversity of citizenship may be brought only in the judicial district where all defendants reside, or in which the claim arose, except as otherwise provided by law.

Where the claim arose in more than one district, "a plaintiff may choose between those two (or conceivably even more) districts that with approximately equal plausibility—in terms of the availability of witnesses, the accessibility of other relevant evidence, and the convenience of the defendant (but not of the plaintiff) — may be assigned as the locus of the claim." *Leroy v. Great Western*, 443 U.S. 173, 185 (1979).

Applying the *Leroy* holding, in a similar case to this, Judge Sofaer, formerly of this district, held that state security employees could bring suit against state officials in the Southern District. *Cheeseman v. Carey*, 485 F. Supp. 203 (S.D.N.Y. 1980). While the state decisions in question had been made in Albany, the court held that the Southern District had a "substantial relationship" to the claim and, thus, the claim "arose" in the Southern District as well as the Northern District. *Id.* at 212. In *Cheeseman* half of the plaintiff class is located in the Southern District, and thus, the court concluded, the challenged practice had been "profoundly felt" in the district.

In this case, the effects of the SED's policy have similarly been profoundly felt in

the Southern District. Female students attending New York City schools are harmed more than elsewhere by the SED's exclusive reliance on SAT scores because they are even less likely to qualify for their scholarships than their female counterparts throughout the state. Moreover, these students take the SAT in the Southern District. Thus, while plaintiffs' claims also arise in the Northern District, they arise in the Southern District as well.

The Southern District is at least an "equally plausible forum." New York City is more accessible for the many expert witnesses who live outside New York. Amici ETS and College Board have offices in the City. It is not overly burdensome for state officials to travel to New York. Defendants have not demonstrated that evidence would somehow be less accessible if this case is maintained in the Southern District.

Finally, the Court concludes that because speed of disposition is important in this case, the interests of justice weigh against a transfer. This Court is familiar with the detailed facts of the case, and substantial proceedings have already occurred before this Court. *See e.g., Cheeseman, supra* 485 F. Supp. at 215. This is not a case where plaintiffs may have chosen their place of venue to harass defendants or to avoid precedents in the Northern District. It appears that plaintiffs merely have chosen a forum that is convenient for the named plaintiffs, teenagers who live in the New York City area.

It is well-established that a plaintiff's choice of forum "is entitled to great weight and will not be disturbed except upon a clear-cut showing that convenience and justice for all parties demand that the litigation proceed elsewhere." *Eastern Refractories v. Forty Eight Insulations* 668 F. Supp. 183, 187 (S.D.N.Y. 1987), *citing Gulf Oil Corp. v. Gilbert*, 330 U.S. 501 (1947). Such a showing has not been made. Accordingly, defendants' venue motion is denied.

B. The Preliminary Injunction

The standard for reviewing a request for a preliminary injunction is well established.

> In this circuit, a preliminary injunction can be granted if plaintiff shows *irreparable injury*, combined with either a *probability of success on the merits*, or a fair ground for litigation and a balance of the hardships in his favor.

The Video Trip Corporation v. Lightning Video, Inc., slip. op. at 1018, (2d Cir. January 20, 1989) (emphasis added), *citing Wainwright Securities, Inc. v. Wall Street Transcript Corp.*, 558 F.2d 91, 94 (2d Cir. 1977), *cert. denied*, 434 U.S. 1014 (1987).

A court need not certify a class prior to granting a preliminary injunction. Defendants improperly rely upon *Hurley v. Ward*, 584 F. 2d 609 (2d Cir. 1978), to support their contention that any injunction must be limited to the individual named plaintiffs. *Hurley* is inapposite because it was brought by an individual plaintiff who did not even seek class certification and members of the purported class were not similarly situated.

Contrary to defendants' argument, courts have consistently granted relief that would have a class-wide effect without first certifying a class. Indeed, in this Circuit, courts have held that where a judgment would run to the benefit not only of the named plaintiffs but also of all others similarly situated, as it would here, class designation is "largely a formality." *Galvan v. Levine*, 490 F. 2d 1255, 1261 (2d Cir. 1973), *cert. denied*, 417 U.S. 936 (1974). *See also Hurley, supra*, 584 F. 2d at 611-612.[31]

Thus, this Court need not certify a class in this case before determining whether plaintiffs have demonstrated the requirements for a preliminary injunction: irreparable harm and a likelihood of success on the merits.[32]

1. Irreparable Harm

Plaintiffs have demonstrated that if the SED is not enjoined from its current practices, they will suffer irreparable harm. Defendants do not dispute that Regents and Empire State Scholarships are prestigious awards, and that students benefit

from receiving such awards. Rather, they merely argue that Regents Scholarships are worth less than Empire State Scholarships, and because it is unlikely that any of the named plaintiffs would receive Empire State awards, plaintiffs have not shown irreparable harm. D. Mem. at 20. This is defendants' standing argument that was dismissed above. To reiterate: first, while named plaintiffs may not receive Empire awards, some members of the putative class would qualify for such awards; second, *all* plaintiffs' chances are reduced by the SED's actions; and third, defendants concede that at least three named plaintiffs will be harmed by the SED's acts. Byrne Aff 10.

When an alleged deprivation of a constitutional right is involved, most courts hold that no further showing of irreparable injury is necessary. *Mitchell v. Cuomo*, 748 F. 2d 804, 806 (2d Cir. 1984). Plaintiffs here go further than merely alleging deprivation of a constitutional right—they document the harm that would result if the SED continued its practiced of reliance upon the SAT.[33] Thus, plaintiffs clearly have demonstrated "irreparable harm."

2. Likelihood of Success on Merits

a. Title IX

Plaintiffs invoke the protections provided by Title IX, which prohibits sex discrimination in federally-funded educational programs.[34] Plaintiffs do not claim that defendants have intentionally discriminated against them based on their sex. Rather, they claim that defendants' practice of sole reliance upon SAT scores to award prestigious state scholarships disparately impacts female students. To this Court's knowledge, this is the first disparate impact case challenging educational testing practices under Title IX.[35]

Neither the Supreme Court nor any court in the Second Circuit has determined whether intent must be shown in Title IX cases.[36] This Court, however, is not without substantial guidance. Recognizing that "Title IX was patterned after Title VI of the Civil Rights Act of 1964," *Grove City College v. Bell* 465 U.S. 555, 556, courts examining Title IX questions have looked to the substantial body of law[37] developed under Title VI, 42 U.S.C. § 2000d, which prohibits race discrimination in federally-funded programs, and Title VII, 42 U.S.C. § 2000e, which prohibits discrimination in employment. *See, e.g., Mabry v. State Board of Community Colleges and Occupational Education*, 813 F.2d 311, 317 (10th Cir.), *cert. denied*, ___U.S.___, 108 S. Ct. 148 (1987); *Haffer v. Temple University*, 678 F. Supp. 517, 539 (E.D. Pa. 1987).

In *Guardians Association v. Civil Service Commission*, 463 U.S. 582 (1983), the Supreme Court held that a violation of Title VI itself requires proof of discriminatory intent. However, a majority also agreed that proof of discriminatory effect suffices to establish liability when a suit is brought to enforce the regulations promulgated under Title VI, rather than the statute itself. *See also Alexander v. Choate*, 469 U.S. 287, 293-294 (1985); *Latinos Unidos de Chelsea v. Secretary of Housing*, 799 F.2d 774, 785 n. 20 (1st Cir. 1986).

Plaintiffs' amended complaint explicitly alleges both violations of Title IX and its implementing regulations. This Court finds no persuasive reason not to apply Title VI's substantive standards to the present Title IX suit. Under analogous circumstances, one district court reasoned:

> The Title IX regulations, like the Title VI regulations at issue in *Guardians*, do not explicitly impose an intent requirement. As there is no reason that a Title IX plaintiff should have a higher burden of proof than a Title VI plaintiff, *see, e.g., Cannon v. University of Chicago*, 441 U.S. 677 (1979) (interpretation of Title IX dependent upon interpretation of Title VI); *Chowdhury v. Reading Hospital & Medical Center*, 677 F. 2d 317 (3d Cir, 1982), *cert. denied*, 463 U.S. 1229 (1983) . . . , I hold that plaintiffs need not prove discriminatory intent to succeed on their claim.

Haffer, 678 F. Supp. at 539-540.

The Title IX implementing regulations, like the regulations promulgated under Title VI, to which Title IX is frequently compared, are consistent with this

interpretation of the comprehensive reach of the statute. Several Title IX regulations specifically prohibit facially neutral policies. For example, the provision governing admissions procedures, 34 CFR 106.21(b) (2), prohibits a recipient from

> administer[ing] or operat[ing] any test or other criteria for admission which has a disproportionately adverse effect on persons on the basis of sex unless the use of such test or criterion is shown to predict validly success in the education program or activity in question and alternative tests or criteria which do not have such a disproportionate adverse effect are shown to be unavailable.

See also 34 C.F.R. §§ 106.22, 106.23 (b), 106.24(d), 106.37(b), 106.52, and 106.53(b).[38]

Based upon a reading of the Title IX regulations, as well as the decisions that apply them, the Court finds that Title IX regulations, like the Title VI regulations at issue in *Guardians*, prohibit testing practices with a discriminatory *effect* on one sex. Consequently, plaintiffs need not prove intentional discrimination.

In Title VII testing cases, the Supreme Court developed a three-pronged formulation to analyze disparate impact claims. Under this scheme, plaintiffs first must show that a facially neutral practice has a disproportionate effect. After such a showing, the burden shifts to defendants to prove a substantial legitimate justification—a "business necessity"—for its practice. The plaintiff then may ultimately prevail by offering either an equally effective alternative practice which has a less discriminatory impact, or proof that the legitimate practices are a pretext for discrimination. *Connecticut v. Teal*, 457 U.S. 440 (1982); *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975); *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); *Sheehan v. Purolator*, 839 F.2d 99, 104 (2d Cir. 1988).[39]

In educational testing cases, instead of requiring defendants to demonstrate a "business necessity," courts have required defendants to show an "educational necessity." For example, the Eleventh Circuit, in *Georgia State Conf. of Branches of NAACP v. State of Georgia*, 775 F.2d 1403 (11th Cir. 1985), held that defendants had a burden of proving that their practices in question bore "a manifest demonstrable relationship to classroom education." *Id.* at 1418. *See also Board of Education v. Harris*, 444 U.S. 130, 151 (1979) ("educational necessity" analogous to "business necessity").

Applying the Title VII formulations to this Title IX case as modified to take into account "educational necessity," this Court finds that plaintiffs have demonstrated a likelihood of success on the merits. Plaintiffs have met their burden of establishing a *prima facie* case through persuasive statistical evidence and credible expert testimony that the composition of scholarship winners tilted decidedly toward males and could not have occurred by a random distribution. *See* Gray Aff. at para. 6; Shapiro T. at 29-30. Defendants have failed to attack plaintiffs' evidence of statewide disparate impact but have instead focused in an *ad hoc* fashion on individual schools and counties. In a case alleging statewide discrimination, such a focus does not rebut plaintiffs' statewide *prima facie* case.

Plaintiffs, moreover, have established that the probability, absent discriminatory causes, that women would consistently score 60 points less on the SAT than men is nearly zero. Gray Aff. at paras. 5-7. Defendants concede that at least half of this differential cannot be explained away by "neutral" variables. Based upon the totality of evidence, then, this Court finds that plaintiffs have demonstrated that the State's practice of sole reliance upon the SAT disparately impacts young women.

Thus, to prevail, defendants must show a manifest relationship between use of the SAT and recognition and award of academic achievement in high school. The Court finds that defendants have failed to show even a reasonable relationship between their practice and their conceded purpose. The SAT was not designed to measure achievement in high school and was never validated for that purpose. Instead, in arguing that the SAT somehow measures high school performance, defendants rely upon anecdotal evidence that the SAT partially tracks what is generally learned in high school math and English courses. This argument is meritless.

Plaintiffs have offered substantial evidence that the SATs do not mirror high school math and English classes. The makers of the SAT describe the test as an "aptitude test"; it does not purport to measure what is learned in classrooms but to predict success in college. The testing format of the SAT measures students' ability to take tests at least as much as it measures substantive material. Defendants' claims that the SAT is an achievenent test are contradicted by defendant Sobol's own recent pronouncement that SAT scores "are a measure of aptitude rather than achievement." *Appeal of Yick Moon Lee* (June 23, 1988), P. Reply Mem. Ex. 4.

Moreover, even if SATs provided a partial measurement of what is learned in high school math and English, these two courses constitute only 20 percent of a high school student's studies. The SAT fails to provide *any* measure of what a student learns in foreign language, science, and social studies courses. Moreover, there can be no serious claim that a test given on one single morning can take into account a student's diligence, creativity and social development and work habits in that student's environment—all part of high school achievement. After a careful review of the evidence, this Court concludes that SAT scores capture a student's academic achievement no more than a student's yearbook photograph captures the full range of her experiences in high school.

Plaintiffs have offered an alternative to sole reliance upon the SAT: a combination of GPAs and SATs. The SED's use of this alternative in 1988 sharply reduced the disparate impact against females caused by the use of the SAT alone. A significantly greater number of female students received scholarships in 1988 than in each prior year in which the SED relied solely upon the SAT. P. Appl I, Ex. 2. Defendants concede that females had a greater opportunity to receive scholarships under the combination system. Defendants also concede that grades are the best measure of high school achievement within the walls of a single school. Instead, they argue that since there is a disparity among schools and their grading systems it is both unfair and impossible to use grades as part of the scholarship eligibility determination. Defendants plan instead to develop a statewide achievement test. While this Court does not dispute the apparent advantages of a statewide achievement test—if indeed a valid test can be developed—it does not agree that pending the implementation of such a test, use of grades would be either unfair or infeasible.

While a combination system—using both GPAs and SATs—is not a perfect alternative, it is the best alternative presently available. The SED is concerned that students in academically superior high schools not be disadvantaged by the use of GPAs. This concern is addressed by the combination system because in effect grades would be weighted by SATs. The SAT component, which cannot properly itself measure achievement, serves to balance the grade component that does. In this way, the SED's concern that use of grades alone will deprive good students in superior high schools of scholarships is ameliorated. Also, as a testing expert explained at the hearing, few students will be displaced if a combination system is used:

> What happens when you add GPA in with the SATs but low grade point averages. And they get replaced by people with slightly lower SATs who have higher—very high grade point averages. So the movement of individuals is not really all that severe, it's ... really just taking scholarships away from the high SAT performers who did not actually achieve in high school ...

Shapiro T. at 36. More importantly, the combination system would be "fair" in the larger sense of the word, because it would better advance the state's goal of awarding high school performance and would better provide *all* students—not just male students or students from selective schools—with an equal opportunity to compete for prestigious state scholarships.

Like its fairness argument, the SED's feasibility argument lacks merit. The SED contends that if it uses GPAs in awarding scholarships, the GPAs will not be in hand until February 24, 1989, the awards process will extend 16 weeks thereafter and thus it will be difficult to inform winners prior to college acceptance dates. The Court rejects this argument. First, based upon evidence detailing the time that is

necessary to process scholarship applications, the Court finds that the awards process can be completed in substantially less than 16 weeks.[40] Second, as of the hearing before the Court on December 21, 1988, the defendants have been on notice that GPAs may be needed and on that date represented to the Court that they had commenced collection of GPAs. The defendants—not the plaintiffs or the Court—selected February 24 as the date grade calculations must be submitted to the SED, and then waited until January 13 to notify schools of that fact. The Court can only assume the SED did so consistent with using GPAs in a timely fashion and, in any event, the SED cannot use its own delay to justify continued reliance upon a discriminatory practice.

The SED cannot justify its discriminatory practice because any alternative would be more difficult to administer. All states giving merit scholarships awards, with the exception of New York and Massachusetts, use GPAs, without concern for either administrative difficulties, grade inflation or the comparability of grades. Lee Aff., P. Reply Mem. Ex. 5; App. I Ex. 14. Any administrative difficulties that the SED experienced in 1988, when it used a combination system, were attributable to the SED's own failure to implement and clarify specific guidelines for the collection of grades, and to provide any enforcement mechanisms to guard against cheating. While the Court, like the *amici* Hewitt School District, does not condone cheating or inaccuracies in grade reporting, it is not the Court's role to police the SED's scholarship program. The Court notes, however, that to verify accuracy, the SED could follow the practice of many states and require school administrators to submit a signed certificate of accuracy.

Faced with a conflict between the SED's administrative concerns on the one hand, and the risk of substantial discriminatory harm to plaintiffs on the other, the Court has little difficulty in concluding that the balance of hardships tips decidedly in plaintiffs' favor. *See Mitchell v. Cuomo*, 748 F. 2d 804, 808 (2d Cir. 1984). The Court finds that plaintiffs have offered a feasible alternative to sole reliance upon SATs. Accordingly, the Court finds that plaintiffs have demonstrated a likelihood of success on the merits of their Title IX claim and, thus, a preliminary injunction is warranted.

b. Equal Protection

Alternatively, a preliminary injunction is warranted because plaintiffs also have established a likelihood that they will succeed on their equal protection claim. The classification of scholarship applicants solely on the basis of SAT scores violates the equal protection clause of the Fourteenth Amendment because this method is not rationally related to the state's goal of rewarding students who have demonstrated academic achievement.

Under the lowest standard of equal protection review—the "rational relationship standard"—"[t]he State may not rely on a classification whose relationship to an asserted goal is so attenuated as to render the distinction arbitrary or irrational." *City of Cleburne v. Cleburne Living Center*, 473 U.S. 432, 446 (1985). Although considerable deference is given to the decisions of legislators and state administrators under the rational basis test, the test "is not a toothless one." *Baccus v. Karger*, 692 F. Supp. 290, 298 (S.D.N.Y. 1988) (invalidating New York bar rule that required applicants for bar admission to have commenced the study of law after their 18th birthday), *citing Schweiker v. Wilson*, 450 U.S. 221, 234 (1980). In a long line of cases, the Supreme Court has applied rational basis scrutiny to strike down legislation where the permissible bounds of rationality were exceeded. *See e.g., Hooper v. Bernalillo County Assessor*, 472 U.S. 612 (1985); *Williams v. Vermont*, 472 U.S. 14 (1985); *Metropolitan Life Insurance v. Ward*, 470 U.S. 869 (1985); *United States Department of Agriculture v. Moreno*, 413 U.S. 528 (1973).

For the reasons stated above, the SED's use of the SAT as a proxy for high school achievement is too unrelated to the legislative purpose of awarding academic achievement in high school to survive even the most minimal scrutiny. The evidence is clear that females score significantly below males on the SAT while they perform equally or slightly better than males in high school. Therefore, the SED's use of the

SAT as the sole criterion for awarding Regents and Empire State Scholarships discriminates against females and, since such a practice is not rationally related to the legislative purpose, it unconstitutionally denies young women equal protection of the laws and must be enjoined on that ground as well.

IV. Conclusion

Defendants' practice of relying solely upon SAT scores in awarding Regents and Empire State Scholarships deprives young women of the opportunity to compete equally for these prestigious scholarships in violation of both Title IX and the Constitution's equal protection clause. Defendants are hereby ordered to discontinue such discriminatory practices and, instead, to award Regents and Empire State Scholarships in a manner that more accurately measures students' high school achievement. For the present year, the best available alternative is a combination of grades and SATs. The SAT component is justified, not as a measure of achievement, but to weight the GPA component. The court, however, does not limit the SED's discretion to develop other alternatives in the future, including a statewide achievement test.

SO ORDERED.

Dated: February 3, 1989
New York, New York

[Signature] John M. Walker
UNITED STATES DISTRICT JUDGE

---

1. The students bring the suit by their parents and next friends. The organizational plaintiffs are the Girls Clubs of America and National Organization for Women.

2. References are made as follows: Amended Complaint ("Am. Complaint"); Affidavit ("Aff."); Testimony of Witness at January 23, 1989 hearing before this Court ("Witness T."); Exhibits ("Ex."); Plaintiffs' Memorandum ("P. Mem."); Plaintiffs' Reply Memorandum ("P. Reply Mem."); Plaintiffs' Appendix ("P. App."); Defendants' Memorandum ("D. Mem.").

3. *Tenth Annual Report of the Education Department* (March 16, 1914), D. Mem., Ex. 3.

4. *Id.* at 30.

5. The college entrance diploma was granted to pupils who pursued courses in approved New York secondary schools and passed the Regents examinations prescribed for such diploma. D. Mem. at 7, n.1.

6. *Tenth Annual Report of the Education Department* at 468 (1914). D. Mem. Ex. 3. Moreover, at that time the college entrance diploma was issued in two very different forms—the college entrance diploma in arts and the college entrance diploma in science. *Id.*

7. *Report of the Select Committee on Higher Education*, State of New York Legislative Document, No. 16, p. 9- (1974), D. Mem. Ex. 4.

8. *Id.* at 29.

9. For example, the most popular Achievement Test is taken by only 25,000 students. Lott T. at 68-69. The SED "felt it would be an imposition on students to require them to pay the additional fee to have to take these College Board achievement tests for scholarship purposes." *Id.*

10. The SED, however, allowed a student to use the ACT as an alternative to the SAT, if the sole test that student had taken was the ACT.

11. The SED has attempted to justify its choice by arguing that some studies found a general correlation between scores obtained on the SAT and scores obtained on the now abandoned Regents Scholarship exams. Lott. T. at 67, 70-71. These studies are discredited, however, because they did not draw any distinctions between men and women, but rather grouped them together in examining the correlation between the tests. Lott. T. at 72. For problems inherent in such a practice, *see infra*. Moreover, like the SAT, the Regents Scholarship examinations were not validated as a measure of high school achievement.

12. Associate of Bureau of Mathematics Education, SED.

13. *See also* Shapiro T. at 44 (Regents questions are more content specific).

14. The State claims that "the primary intent of the legislation is to include a 'measure of high school performance' in the scholarship eligibility criteria because it was believed that high school performance is a better predictor of college performance than are SAT scores." D. Mem. at 10. In support of its claim, the State cites one isolated exchange from the Hearing on the Implementation of Changes in Criteria for Awarding Regents Scholarships and Empire State Scholarships of Excellence, New York State Senate and Assembly Standing Committee on Higher Education, May 25, 1988, pp. 19-21, P. App. I, attachment 4 (hereafter "May Hearing").

15. The GPA calculation was to be based upon three years of English, three years of social studies, two years of math, one year of science, and any three-year sequence in the student's major. Lott T. at 103. However, the SED issued no guidelines as to how to compute average GPAs. Kenneth Ormiston, SED Bureau Chief, told one school: "when calculating grade point average, you may use either weighted or unweighted grades." Letter of Nov. 23, 1987, P. App. I, Ex. 11. After receiving "hundreds of phone calls" about computation of grade point averages, one SED official, Jim Brown, wrote to Ormiston, urging that the policy be clarified. Brown Memo at 1, App. I, Ex. 12. However, SED officials did not issue clarifying instructions. *See* Hearings at 82-85.

16. The Attorney for the SED addressed the Commissioner's apparent about face at trial by explaining, "[a]t the time the statement was made, the statute had not yet expired. And he was assuming that if the statute was extended it would be made on a grade point average." T. at 199.

17. "Validity refers to the degree to which evidence supports the inference and use of test scores." Tittle Aff. para. 11.

18. Emery T. at 47. See also Tittle Aff. at paras. 6, 10-19, 29; Campbell Aff. at paras. 6, 12-17, 20; Behnke Aff. at paras. 2-7. When the regression equation is based on what is called a common regression line (with males and females together), a male and female with the same SAT scores will obtain different grade point averages, with the female's actual grade point average being somewhat higher than her predicted average. The male's actual grade point average will be somewhat lower than that predicted. Campbell Aff. at paras. 11-14.

19. *See* 1988 Profile of SAT and Achievement Test Takers - National Report p. iii (ETS, 1988). P. App. I Ex. 7. The ETS argues only that "it would be incorrect to suggest . . . that the College Board and ETS guidelines are the product of any conclusion that the SAT is biased in any way . . . " ETS and College Board *Amici* Brief at 9.

20. Statement by NACAC on the role of Standardized Testing in the College Admission Process, P. App. II, Part 3, Ex. X; Richard Stewart, vice-president of committee for admissions practices, NACAC, T. at 58; NACAC Statement of Principles of Good Practices (December 1988); Burnett Aff. at para. 3.

21. Indeed, without a prescribed curriculum (as in New York State), it would be very difficult to prove the content validity of the SAT. Shapiro T. at 42.

22. The SED's claim that the SAT is an achievement test contradicts the position taken by Commissioner Sobol himself, on June 23, 1988 in the *Appeal of Yick Moon Lee*. P. Reply Mem., Ex. 4. In that case, Sobol stated that SAT scores "are a measure of aptitude rather than achievement." (In *Yick Moon Lee*, a student appealed to the Commissioner to enjoin the practice in his school of using a combination of SAT and GPA to calculate class rank. Commissioner

Sobol ordered that the practice be discontinued because SAT scores do not measure "actual student achievement.")

23. *1988 Profile of SAT Takers*, The College Board, P. App. I Ex. 7, p. iii. These undisputed results are summarized in Appendix B, *infra*.

24. It is debatable whether all of these factors are indeed "neutral" and do not to some degree reflect systemic sex discrimination.

25. Shapiro T. at 50; Willingham Aff., Ex. 6. *See also* Clark and Grandy at 18 P. App. II, Part 2, Ex. D. and Gamache and Novick, P. App. II, Part 3, Ex. M.

26. According to the SED's own estimates of the 1988-89 competition, 56 percent of the winners of the Regents Scholarship will be male if only SAT results are used to determine scholarship winners despite the fact that 53 percent of all the competitors are female. Statistical review of the Awarding of the 1988 New York State Scholarships (April 1988) p.l App. I, Ex. 2.

27. As statistical significance is generally recognized to be .05 standard deviations from the mean, there is no doubt that these figures are statistically significant. The reason why the Regents results are more significant than the Empire State results is that sample size greatly affects calculations of standard deviations, and 25,000 Regents awards are given annually as compared to only 1,000 Empire State awards.

28. *See* Plaintiff's Affidavits, App. II: Bonzon at para. 4; Hart at para. 3; Lewis at para. 5; Sharif at paras. 3, 4; Sultan at para. 2; Greenblatt at 4, 5; Taylor at para. 8.

29. Defendants argue that plaintiffs do not have standing because some female students' chances of winning scholarships will be reduced if a combination of grades and SATs are used to determine the awards. D. Mem. at 23. This is irrelevant. The fact is that as a group women's chances are improved. Plaintiffs would rarely succeed in educational testing cases if courts accepted defendant's argument, because changes in *any* test have differing effects on a broad class of plaintiffs.

30. The state awards 25,000 Regents scholarships of $250 and 1000 Empire State Scholarships of $2,000 to the top Regents Scholarship winners.

31. Moreover, the Supreme Court has held in connection with statute of limitations questions that class-wide relief is appropriate unless and until the class is dismissed. *Crown, Clark and Seal Co. v. Parker*, 462 U.S. 345 (1983); *American Pipe and Construction Co. v. Utah*, 414 U.S. 538, 551 (1974).

32. While plaintiffs filed a motion for class certification on January 30, 1989, the Court will not consider this motion until defendants have filed their response.

33. See P. Mem. at 20-23; Bonzon Aff. at para. 4; Hart Aff. at para. 4; Lewis Aff. at para. 5; Mackenzie Aff. at para. 3; Sharif Aff. at paras. 3,4; Sultan Aff. at para. 2. Defendants have not disputed these affidavits.

34. Title IX provides, in pertinent part:

> (a) No person in the United States shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving any federal assistance.
>
> 20 U.S.C. § 1681 (a). Recently, Congress broadened the scope of Title IX so that it applies institution-wide. Civil Rights Restoration Act, 20 U.S.C. § 1687 (became law on March 22, 1988). This Act directly reversed the Supreme Court's decision in *Grove City College v. Bell*, 465 U.S. 555 (1984), which had limited the coverage of Title IX to specific programs or activities which actually receive federal funds.

35. The most common Title IX cases challenge regulations that prohibit female students from participating in high school sports. In general, courts have had little difficulty in concluding that such regulations deny female students equal protection of the laws. *See, e.g., Brenden v. Independent School District*, 477 F. 2d 1292 (8th Cir. 1973); *Morris v. Michigan State Board of Education*, 472 F.2d 1207 (6th Cir. 1973); *Hoover v. Meiklejohn*, 430 F. Supp. 164 (D. Col. 1977).

36. Judge Sweet of this district applied a disparate impact analysis under Title IX in *Fulani v. League of Women Voters Education Fund*, 684 F. Supp. 1185 (S.D.N.Y. 1988). However, he carefully *assumed* that a disparate impact is appropriate under Title IX without actually *deciding* that such was the case. *Id.* at 1193.

37. Many of these cases have challenged teacher competency tests as being racially discriminatory. *See generally* Rebell, *Disparate Impact of Teacher Competency Testing on Minorities: Don't Blame the Test-Takers—or the Tests*, 4 YALE L. & POL. REV. 375 (1986).

38. The Tenth Circuit made a similar observation in *Mabry*, 813 F.2d at 316-17 n.6.

39. Defendants' mistakenly rely upon Justice O'Connor's plurality opinion in *Watson v. Fort Worth Bank & Trust*, 108 S. Ct. 2777, 1790 (1988), to argue that the Supreme Court now requires a greater quantum of proof in disparate impact cases. The portion of Justice O'Connor's opinion containing the alleged change in law was only joined by three other members of the Court. Thus, it is not law. The Court's holding in *Watson*, that subjective employment practices can be challenged under disparate impact analysis, does not affect the outcome of this case.

40. While the Court will not play the role of administrator and detail the time saving techniques that the SED could use, the Court does note that the SED could save considerable time if it notifies students of their awards by listing names in a New York newspaper, a procedure used successfully by the New York Board of Bar Examiners.